НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 519.816

До захисту допущено
В. о. завідувача кафедри ММСА
О.Л.Тимощук
«___» _____ 2020 р.

Магістерська дисертація

на здобуття ступеня магістра за спеціальністю 124 Системний аналіз
на тему: «Система аналізу та категоризації еколого-економічних даних для прогнозування розвитку територіальних громад з використанням SAS технологій»

Виконала:
студентка ІІ курсу, групи КА-92 мп
Дітковська Юлія Василівна _____

Керівник:
доцент кафедри ММСА, к.е.н.
Просянкіна-Жарова Т. І. _____

Рецензент:
старший науковий співробітник
Інституту телекомунікацій і глобального
інформаційного простору НАН України,
к.т.н., доц.
Терентьєв О.М. _____

Засвідчую, що у цій магістерській дисертації
немає запозичень з праць інших авторів
без відповідних посилань
Студент _____

Київ
2020

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність — 124 «Системний аналіз»

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри ММСА
О. Л. Тимощук
«___» _____ 2020 р.

ЗАВДАННЯ
на магістерську дисертацію студенту Дітковській Юлії Василівні

1. Тема дисертації: «Система аналізу та категоризації еколого-економічних даних для прогнозування розвитку територіальних громад з використанням SAS технологій», науковий керівник дисертації Просянкіна-Жарова Тетяна Іванівна, к.е.н., доцент, затверджені наказом по університету від 02 листопада 2020 р. № 3182-с

2. Термін подання студентом дисертації: 14 грудня 2020 р.

3. Об'єкт дослідження: новини, статті та документи, пов'язані з еколого-економічними даними Украни.

4. Предмет дослідження: методи тестової аналітики, інструменти SAS для інтелектуальної обробки неструктурованих даних.

5. Перелік завдань, які потрібно розробити:

1. Огляд предметної області та аналіз існуючих систем для інтелектуального аналізу даних;
2. Збір даних для аналізу;
3. Розробка програмного комплексу, що забезпечуватиме використання існуючих та розроблених методів для аналізу та категоризації еколого-економічних даних.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

1). Схема архітектури розробленої системи

2). Приклади функціонування створеного програмного продукту

3). Таблиці у розділі стартап-проекту

7. Дата видачі завдання: 01 вересня 2020 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації
1.	Концептуальний вступ дисертації. Формулювання об’єкта, предмета, цілі, завдань, новизни, практичної значущості результатів	01.09.2020—20.09.2020
2.	Перший розділ. Огляд літературно-інформаційних джерел.	21.09.2020—30.09.2020
3.	Другий розділ. Опис методів текстової аналітики	01.10.2020—20.10.2020
4.	Третій розділ. Огляд програмного продукту та аналіз отриманих результатів	21.10.2020—16.11.2020
6.	Четвертий розділ. Стартап-проект	17.11.2020—20.11.2020
7.	Оформлення роботи	21.11.2020—01.12.2020

Студентка Ю.В. Дітковська

Науковий керівник дисертації Т.І. Просянкіна-Жарова

РЕФЕРАТ

Магістерська дисертація: 84 с., 27 рис., 28 табл., 1 додаток, 26 джерел.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, КАТЕГОРИЗАЦІЯ, КЛАСТЕРИЗАЦІЯ, SAS TEXT MINER, ЛАТЕНТНО-СЕМАНТИЧНИЙ АНАЛІЗ, СИНГУЛЯРНИЙ РОЗКЛАД.

Тема магістерської дисертації «Система аналізу та категоризації еколого-економічних даних для прогнозування розвитку територіальних громад з використанням SAS технологій».

Актуальність магістерської дисертації обумовлена динамічним зростанням кількості інформації, яка більшою мірою знаходиться у неструктурованому вигляді, і самотужки опрацювати її неможливо. Інтелектуальний аналіз даних дозволяє досить швидко виявити факти, взаємозв'язки та твердження, які в протилежному випадку залишились би непоміченими у неструктурованих текстових даних.

Об'єкт дослідження - новини, статті та документи, пов'язані з еколого-економічними даними Украни.

Предмет дослідження - методи тестової аналітики, інструменти SAS для інтелектуальної обробки неструктурованих даних.

Метою магістерської дисертації є аналіз та категоризація еколого-економічних даних для досить швидкого реагування на відповідні зміни у сферах екології і економіки.

Результати роботи: розроблено прикладну систему аналізу еколого-економічних даних за допомогою SAS Enterprise Miner.

Новизна роботи: після обробки інформації користувач отримує не тільки результати категоризації, а також аналіз зв'язків між ключовими термінами.

ABSTRACT

Master's dissertation: 84 p., 26 fig., 28 tables, 1 application, 26 sources.

DATA ANALYSIS, CATEGORIZATION, CLUSTERIZATION, SAS TEXT MINER, LATENT SEMANTIC ANALYSIS, SINGULAR VALUE DECOMPOSITION.

The theme of my master's dissertation "System of analysis and categorization of ecological and economic data for forecasting the development of territorial communities using SAS technologies ".

The relevance of the Master's dissertation is due to the massive increase of information in the world, which is lot of them in a structured form, and it is impossible to process it alone. Data mining allows you to quickly identify facts by interacting with connections and statements that would otherwise remain unknown in structured text data.

Object of research - news, articles and documents presented with ecological and economic data of Ukraine.

The subject of research - methods of test analytics, SAS tools for intelligent processing of unstructured data.

The purpose of the master's dissertation is the analysis and categorization of ecological and economic data for rapid response to relevant changes in ecology and economics individual territorial communities.

Results of the research: the applied system of the analysis of ecological and economic data with SAS Enterprise Miner.

The novelty of the work: after processing the information, the user receives not only the results of categorization, but also the analysis of relationships between key terms.

ЗМІСТ

ВСТУП

У сучасному світі існує безліч джерел нової інформації, і самотужки опрацювати їх неможливо. Оскільки більша частина інформації у інтернеті, новинах, текстових документах знаходиться у неструктурованому вигляді, цю інформацію треба попередньо обробити, щоб використовувати у програмних додатках. Саме тому виникає необхідність у текстовому аналізі даних. Аналіз тексту виявляє факти, взаємозв'язки та твердження, які в протилежному випадку залишились би непоміченими у неструктурованих текстових даних.

Територіальні громади є складними соціально-еколого-економічними системами. Такі системи характеризуються значними обсягами різнорідної та неструктурованої, неоднозначної та неповної інформації, тому такі системи потребують опрацювання величезної кількості інформації і досить швидкого реагування на відповідні зміни. Тому вирішено створити інформаційно-аналітичну систему, призначену для автоматизованої підтримки прийняття рішень, яка узагальнює знання про еколого-економічні данні. Після опрацювання даних, ця інформація перетворюється у структуровану форму, яка може бути далі проаналізована експертами для пошуку можливих альтернатив для розвитку громад та для побудови можливих сценаріїв.

Для вирішення цієї задачі, було використано програмні засоби SAS, оскільки компанія SAS пропонує потужний набір інструментів, призначених для текстової аналітики, інтелектуального аналізу даних, прогнозного моделювання, такі як SAS Crawler, SAS Search, Indexing, SAS Enterprise Content Categorization, SAS Ontology Management, SAS Sentiment Analysis Studio, SAS Text Miner.

РОЗДІЛ 1
ПРОБЛЕМИ ВИКОРИСТАННЯ ЕКОЛОГО-ЕКОНОМІЧНИХ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ РОЗВИТКУ ТЕРИТОРІАЛЬНИХ ГРОМАД

1.1 Використання інтелектуального аналізу даних для дослідження стану та перспектив розвитку екологічних, соціальних, економічних систем для розв’язання задач урядування

Складність створення аналітичного інструментарію для розв’язання задач підтримки прийняття рішень в галузі урядування перш за все пов’язана з необхідністю урахування значної кількості кількісних та якісних факторів. Зокрема, особливостей життєвого циклу різних рівнів соціально-економічної системи держави, її структуру, зв'язки, суспільно-політичні чинники, екологічні фактори, зовнішні та внутрішні впливи різного характеру, а також передбачити можливі сценарії розвитку подій, наслідки та потенційні ризики помилок під час прийняття управлінських рішень. Крім того, необхідно забезпечити сумісність інформаційних ресурсів органів державного управління різних рівнів, бізнесових структур [1].

Територіальні громади є складними соціально-еколого-економічними системами. Такі системи характеризуються адаптивністю, схильністю до самоорганізації, а процеси, що відбуваються в них – значною нестаціонарністю, нелінійністю, складністю, наявністю невизначеностей структурного і параметричного типів, значними обсягами інформації, що стосується різних предметних областей, але досить часто недостатністю такої, що характеризує змінні і параметри систем належним чином для побудови адекватних прогнозуючих та управлінських математичних моделей [2]. Складність створення відповідного аналітичного інструментарію пов’язана з необхідністю урахування значної кількості факторів, зокрема особливостей життєвого циклу конкретної соціально-еколого-економічної системи, її структури, зв'язків, суспільно-політичних чинників, зовнішніх та внутрішніх впливів різного

характеру, а також передбачити проблемні ситуації, сценарії розвитку подій, можливі наслідки та потенційні ризики.

У зв’язку із цим виникає необхідність створення таких сучасних технологій прийняття рішень, які давали б змогу в умовах обробки значних обсягів різнорідної та неструктурованої, а в окремих випадках, обмеженої та неповної або і надлишкової інформації, приймати коректні рішення щодо стратегії соціально-економічного розвитку, запобіганню загроз національній безпеці, зменшенню ризиків різних типів, а також забезпечували б зворотній зв'язок щодо ефективності прийнятих оперативних та стратегічних управлінських рішень. Все це свідчить про високу актуальність дослідження.

Особливість проведення дослідження полягає в зборі накопичених знань фахівців з означеної проблеми, що обробляються із використанням сучасних високопродуктивних технологій накопичення та доступу даних Big Data, на основі яких із використанням методів Data Science виконується повний цикл задач аналітичного процесу, починаючи від аналізу якості даних та роботи з пропусками, до побудови сценаріїв з використанням моделей-кандидатів розвитку суб’єкта дослідження на основі методів інтелектуального аналізу даних.

Ідея проекту полягає у створенні інформаційно-аналітичної системи, призначеної для автоматизованої підтримки прийняття рішень, яка буде вирішувати проблеми дослідження сучасного стану регіону чи конкретної адміністративно-територіальної одиниці, консолідувати найбільш вагомі показники, за якими можна оцінити реалізацію реформ, запропонованих у програмі Уряду та прийняти рішення щодо оперативного та стратегічного управління. Запропонована інформаційно-аналітична система допоможе зменшити час для опрацювання величезної кількості текстових документів для більш швидкого реагування на виникненні проблеми.

Методологія моделювання та прогнозування стану та розвитку екологічних, економічних та соціальних систем методами інтелектуального аналізу даних може бути використана в органах державного та регіонального

управління, місцевого самоврядування, бізнесових структурах не лише для дослідження результатів проведення реформ, а й з метою оцінювання роботи їх менеджменту, виконання інвестиційних програм та стратегій, використання грантів, тощо, за потреби, може бути використана і для розробки бізнес-планів та інвестиційних програм, регіональних стратегій, програм соціально-економічного розвитку.

Дослідивши досвід вітчизняних та закордонних фахівців у галузі системного аналізу функціонування складних об'єктів та систем, математичного моделювання і прогнозування, використання методів інтелектуального аналізу даних, проектування та реалізації інформаційних систем, слід зазначити, що питання створення комп'ютерних інформаційних систем з розвинутим інструментарієм моделювання та прогнозування залишається недостатньо опрацьованим.

Більшість існуючих програмних продуктів не можуть бути використані для розв'язування складних інтелектуальних задач, як з невизначеністю і ризиком, так і необхідністю залучення даних і знань з різних предметних областей. Це пов'язано з високою ціною, недостатньою функціональністю, обмеженістю набору реалізованих моделей, недостатністю деталізації та відсутністю урахування тенденцій щодо змін у національній економіці, в довкіллі тощо. В частині моделювання процесів обробки та розподілу дискретних потоків в комунікаційних мережах у відомих підходах зазвичай використовуються лінійні математичні моделі, що добре зарекомендували себе на практиці, різновиди симплекс-методу, методи еліпсоїдів, внутрішніх точок та декомпозиції за змінними і обмеженнями задачі. Існують і універсальні програмні засоби розв'язання задач математичного програмування, такі як LINGO, CPLEX, SOPLEX, MINOS та ін. Однак в цих засобах не враховуються специфічні особливості окремих задач, які призводять до неможливості їх безпосереднього застосування.

Слід зазначити, що комп'ютерні інформаційні системи, використовувані у інформаційно-аналітичній діяльності органів державного управління,

зазвичай не містять розвинених підсистем моделювання та прогнозування нелінійних нестаціонарних процесів у соціально-економічних системах різних рівнів, а також достатніх засобів підтримки прийняття рішень. Також не існує достатньо ефективних систем підтримки прийняття рішень для розв’язання задач децентралізації та регіонального розвитку.

Більшість з програмних засобів призначена для управління розвитком бізнесових структур. Прикладами таких систем можна назвати продукти таких провідних світових компаній як Oracle, SAS Institute, SAP, IDC, тощо. Відомі в Україні й розробки корпорацій «Парус», «1С», Прогноз», «Галактика», підприємств ДП "ГОЛОВФІНТЕХ" та ДП "Держінформресурс", тощо. Однак, враховуючи потреби сектору державного управління, а особливо його регіонального рівня та громад у якісних аналітичних продуктах інтелектуально-комп’ютерного спрямування, слід зазначити, що цей сегмент ІТ-ринку залишається недостатньо опрацьованим.

1.2 Оосбливості застосування інтелектуального тексту для вирішення задач підвищення якості аналітичної діяльності

Інтелектуальний аналіз тексту - це технологія штучного інтелекту, яка використовує обробку природної мови для перетворення вільного формату (неструктурованого) тексту в документах та базах даних у нормалізовані, структуровані дані, придатні для аналізу або алгоритмів машинного навчання.

Інтелектуальний аналіз тексту, представляє собою процес вивчення великих колекцій документів для виявлення нової інформації або допомоги в справах з конкретних дослідницьких питань [3].

Після інтелектуального аналізу тексту, ця інформація перетворюється у структуровану форму, яка може бути далі проаналізована або представлена безпосередньо за допомогою кластерних таблиць HTML, інтелектуальних карт,

діаграм і т. д. В інтелектуальному аналізі тексту використовуються різні методології для обробки тексту, однією з найбільш важливих з яких є обробка природної мови.

Структуровані дані, створені за допомогою інтелектуального аналізу тексту, можуть бути інтегрованими в базу даних, сховища даних або інформаційні панелі бізнес-аналітики та використовуватися для опису, або прогнозування.

Розуміння природної мови допомагає машинам «читати» текст, моделюючи здатність людини розуміти природну мову, наприклад англійську, іспанську або китайську. Обробка природної мови включає в себе як розуміння природної мови, так і генерацію природної мови, що імітує здатність людини створювати текст, наприклад, щоб узагальнити інформацію або взяти участь в діалозі.

Як технологія, обробка природної мови досягла повноліття за останні десять років, коли такі продукти, як Siri, Alexa і голосовий пошук Google, використовують обробку природної мови, щоб розуміти запити користувачів і відповідати на них. Складні додатки для інтелектуального аналізу тексту також були розроблені в таких різних областях, як медичні дослідження, управління ризиками, обслуговування клієнтів, страхування (виявлення шахрайства) і контекстна реклама [3].

Сучасні системи обробки природної мови можуть аналізувати необмежені обсяги текстових даних без втоми і послідовно і неупереджено. Вони можуть розуміти концепції в складних контекстах і розшифровувати двозначність мови, щоб витягувати ключові факти та взаємозв'язки. З огляду на величезну кількість неструктурованих даних, які створюються кожен день, від електронних медичних карт до повідомлень в соціальних мережах, ця форма автоматизації стала критично важливою для ефективного аналізу текстових даних.

Машинне навчання - це технологія штучного інтелекту, яка дає системам можливість автоматично вчитися на власному досвіді без необхідності явного

програмування і може допомогти вирішувати складні проблеми з необхідною для людей точністю [3].

Однак машинне навчання вимагає добре продуманих вихідних даних для навчання, які зазвичай недоступні з таких джерел, як електронні медичні картки, соціальні мережі або наукова література, де велика частина даних складається з неструктурованого тексту.

Обробка природної мови може витягувати чисті, структуровані дані, необхідні для управління просунутими прогностичними моделями, що використовуються в машинному навчанні, тим самим знижуючи потребу в дорогих ручних анотаціях навчальних даних [3].

У той час як традиційні пошукові системи, такі як Google, тепер пропонують уточнення, такі як синоніми, автозаповнення і семантичний пошук (історія і контекст), переважна більшість результатів пошуку вказують тільки на розташування документів, тому необхідно витрачати свій час, щоб знайти необхідні дані шляхом прочитання окремих документів.

Обмеження традиційного пошуку посилюються зростанням великих даних за останнє десятиліття, що допомогло збільшити кількість результатів, що повертаються на один запит такої пошукової системою, як Google, з десятків тисяч до сотень мільйонів [3].

Зі зростанням обсягів текстових великих даних використання технологій штучного інтелекту, таких як обробка природної мови і машинне навчання, стає ще більш необхідне.

Онтології, словники - потужні інструменти, що допомагають в пошуку, отриманні даних і інтеграції даних. Вони є ключовим компонентом багатьох інструментів інтелектуального аналізу тексту і надають списки ключових понять з іменами і синонімами, часто розташованими в ієрархії [3].

Пошукові системи, інструменти текстової аналітики і рішення для обробки природної мови стають ще більш потужними з онтологіями для конкретних предметних областей. Онтології дозволяють зрозуміти реальне значення тексту, навіть якщо воно виражається по-різному (наприклад,

тайленол проти ацетамінофену). Методи штучного інтелекту розширюють можливості онтологій, наприклад, дозволяючи порівняти терміни з різним написанням і беручи до уваги контекст [3].

Специфікація онтології включає словник термінів і формальні обмеження на її використання. Для корпоративної обробки природної мови потрібно декілька словників, онтологій і пов'язаних стратегій для визначення концепцій в їх правильному контексті:

Підходи на основі шаблонів для таких категорій, як вимірювання, мутації і хімічні назви, які можуть включати нові (невидимі) терміни;

Ідентифікація, анотації і перетворення понять на основі правил і специфіки предметної області;

Інтеграція словників клієнтів для створення індивідуальних анотацій;

Розширений пошук для визначення діапазонів даних по датах, числовим значенням, площі, концентрації, процентному змісту, тривалості, довжині і вазі.

Для ефективної обробки природної мови потрібні деякі функції, які описані нижче.

Існує величезна різноманітність документів і текстового контексту, включаючи джерела, формат, мову і граматику. Для боротьби з цим розмаїттям потрібно ряд методологій:

- перетворення внутрішніх і зовнішніх форматів документів (наприклад, HTML, Word, PowerPoint, Excel, текст PDF, зображення PDF) в стандартизований формат з можливістю пошуку;
- можливість ідентифікувати, позначати теги і виконувати пошук в певних розділах (областях) документа, наприклад: зосередження пошуку для видалення шуму з довідкового розділу статті;
- лінгвістична обробка для визначення значущих одиниць в тексті, таких як речення, групи іменників і дієслів, разом зі взаємозв’язками між ними;
- семантичні інструменти, які ідентифікують концепції в тексті, такі як громади і території, і нормалізують до концепцій зі стандартних онтологій;

- розпізнавання образів для виявлення та ідентифікації категорій інформації, які нелегко визначити за допомогою словникового підходу. До них відносяться дати, числова інформація, терміни;
- можливість обробки вбудованих таблиць в тексті, відформатованих з використанням HTML або XML, або у вигляді довільного тексту.

Робота по аналітиці тексту зазвичай складається з наступних компонентів [4]:

- Зниження розмірності - важливий метод попередньої обробки даних. Метод використовується для визначення кореневого слова для реальних слів і зменшення розміру текстових даних.
- Пошук інформації - це підготовчий етап: збір або ідентифікація набору текстових матеріалів в Інтернеті або в файлової системі, базі даних чи в іншому сховищу даних [4].
- Хоча деякі системи текстової аналітики застосовують виключно передові статистичні методи, багато інших застосовують ширшу обробку природної мови, наприклад, синтаксичний аналіз або інші типи лінгвістичного аналізу.
- Розпізнавання іменованих сутностей - це використання географічних довідників або статистичних методів для ідентифікації іменованих текстових елементів: людей, організацій, географічних назв, символів біржових котирувань, визначених скорочень [4].
- Усунення неоднозначності - використання контекстних підказок - може знадобитися, щоб вирішити, де, наприклад, "Форд" може відноситися до колишнього президента США, виробнику автомобілів, кінозірці, або будь-якої іншої сутності.
- Розпізнавання об'єктів, що ідентифікуються за шаблоном: такі функції, як номери телефонів, адреси електронної пошти, кількості (із зазначенням одиниць виміру), можна розпізнати за допомогою регулярних виразів або інших збігів з шаблоном [4].

- Кластеризація документів: ідентифікація наборів схожих текстових документів, ідентифікація словосполучень та інших термінів, що відносяться до одного і того ж об'єкту.
- Отримання взаємозв'язків, фактів і подій: ідентифікація асоціацій між сутностями та іншою інформацією в тексті
- Аналіз настроїв включає в себе розпізнавання суб'єктивного (на відміну від фактичного) матеріалу і отримання різних форм інформації про поведінку: думки, настрої і емоції. Методи текстової аналітики корисні при аналізі настроїв на рівні сутності, концепції або теми.
- Кількісний аналіз тексту - це набір методів, звернених до соціальних наук, де або людина-суддя, або комп'ютер витягають семантичні або граматичні відносини між словами, щоб з'ясувати значення або стилістичні патерни, як правило, випадкового особистого тексту з метою психологічного профілювання і ін [4].

Існує безліч сфер застосування інтелектуального аналізу тексту.

Промислове виробництво створює ідеальні умови для застосування технологій інтелектуального аналізу даних. Причина - в природі технологічного процесу, який повинен бути відтвореним і контрольованим. Всі відхилення протягом процесу, що впливають на якість вихідного результату, також знаходяться в заздалегідь відомих межах. Таким чином, створюється статистична стабільність, першорядну важливість якої відзначають в роботах по класифікації. Досвід роботи компаній, що пропонують рішення Data Mining для промислового виробництва, також свідчить про успішність такої інтеграції. Прикладом застосування Data Mining в промисловості може бути прогнозування якості виробу в залежності від вимірюваних параметрів технологічного процесу [5].

Класичним прикладом застосування Data Mining на практиці є вирішення проблеми про можливу некредитоспроможність клієнтів банку. Це питання, що тривожить будь-якого співробітника кредитного відділу банку, можна вирішити і інтуїтивно. Якщо образ клієнта в свідомості банківського службовця

відповідає його уявленню про кредитоспроможність клієнта, то кредит видавати можна, інакше - відмовити. За схожою схемою, але більш продуктивно і повністю автоматично працюють встановлені в тисячах американських банків системи підтримки прийняття рішень з вбудованою функціональністю Data Mining. Позбавлені суб'єктивної упередженості, вони спираються у своїй роботі тільки на історичну базу даних банку, де записується детальна інформація про кожного клієнта і, в кінцевому підсумку, факт його кредитоспроможності. Класифікаційні алгоритми Data Mining обробляють ці дані, а отримані результати використовуються далі для прийняття рішень [5].

Аналіз кредитного ризику полягає, перш за все, в оцінці кредитоспроможності позичальника. Це завдання вирішується на основі аналізу накопиченої інформації, тобто кредитної історії попередніх клієнтів. За допомогою інструментів Data Mining (дерева рішень, кластерний аналіз, нейронні мережі та ін.) банк може отримати профілі сумлінних і неблагонадійних позичальників. Крім того, можна класифікувати позичальника за групами ризику, а значить, не тільки вирішити питання про можливість кредитування, але і встановити ліміт кредиту, відсотки по ньому і термін повернення [5].

Шахрайство з кредитними картками є серйозною проблемою, тому що збитки від цього вимірюються мільйонами доларів щорічно, а зростання кількості шахрайських операцій становить, за оцінками експертів, від 15 до 25% щорічно. У боротьбі з шахрайством технологія Data Mining використовує стереотипи підозрілих операцій, створені в результаті аналізу величезної кількості транзакцій - як законних, так і неправомірних. Досліджується не тільки окремо взята операція, але і сукупність послідовних в часі транзакцій. Крім того, алгоритми і моделі (наприклад, нейронні мережі), наявні в складі продуктів Data Mining, здатні тестуватися і самонавчатися. При спробі здійснення підозрілої операції засоби інтелектуального аналізу даних оперативно видають попередження про це, що дозволяє банку запобігти

незаконним діям, а не усувати їх наслідки. Використання технології Data Mining дозволяє скоротити кількість порушень на 20-30% [5].

Галузі наук про життя і охорону здоров'я генерують велику кількість текстових та числових даних. У медицині це інформація про пацієнтів, захворювання, ліки, симптоми і методи лікування захворювань і багато іншого. У екології - інформація про забруднення повітря чи води, зміну клімату, хімічний аналіз продуктів харчування. Відфільтрувати відповідний і релевантний текст для прийняття рішення з великого біологічного сховища - велика проблема. Ці дані містять різний характер даних, складну, довгу і технічну лексику, що дуже ускладнює процес. Інструменти інтелектуального аналізу тексту в екологічній області дають можливість отримувати цінну інформацію, і робити висновки про взаємозв'язок між різними видами живих організмів та їх впливом на оточуючу середу. Використання відповідних інструментів інтелектуального аналізу тексту в медичній сфері допомагає оцінити ефективність медичних методів лікування, які демонструють ефективність, шляхом порівняння різних захворювань, симптомів та курсу їх лікування. Інтелектуальний аналіз тексту використовується при виявленні біомаркерів, в фармацевтичній промисловості, клінічному аналізі торгівлі, у доклінічних дослідженнях безпечності препаратів, дослідженні генів хвороб. екології, плануванні територій [6].

Пакети програмного забезпечення для аналізу тексту доступні для аналізу додатків соціальних мереж, щоб відстежувати і аналізувати відкритий текст з інтернет-новин, блогів, електронної пошти і т. д. Інструменти інтелектуального аналізу тексту допомагають ідентифікувати і аналізувати кількість постів, лайків та підписників в соціальних мережах. Такий аналіз показує, як люди реагують на різні пости, новини і як вони поширюються. Він показує поведінку людей, що належать до певної вікової групи або певних спільнот, та демонструє відмінності в поглядах на один і той же пост [7,8].

Інтелектуальний аналіз тексту відіграє важливу роль в бізнес-аналітиці, допомагаючи організаціям і підприємствам аналізувати своїх клієнтів і

конкурентів для прийняття більш ефективних рішень. Він дає більш глибоке уявлення про бізнес і дає інформацію про те, як підвищити задоволеність клієнтів і отримати конкурентні переваги. Інструменти інтелектуального аналізу тексту допомагають приймати рішення про організації і приймати коригуючі заходи, генеруючи попередження о хорошій чи поганій продуктивності, зміні ринку. Ці інструменти також допомагають в телекомунікаційній галузі, комерційних додатках і системі управління взаємовідносин з клієнтами[6].

1.3 Етапи інтелектуального аналізу даних

Особливістю дослідження є необхідність збору та обробки значних обсягів накопиченої інформації з означеної проблеми,. для вирішення даної задачі пропонується використовувати сучасні методи інтелектуального аналізу даних. Етапи інтелектуального аналізу представлено на рисунку 1.1.

Рисунок 1.1 – Етапи інтелектуального аналізу даних

На першому етапі виконується осмислення поставленої задачі і уточнення цілей, які повинні бути досягнуті методами Data Mining. Важливо правильно

сформулювати цілі і вибрати необхідні для їх досягнення методи, тому що від цього залежить подальша ефективність всього процесу [5].

Другий етап полягає у приведенні даних до форми, придатної для застосування конкретних методів Data Mining. Вид перетворень, що здійснюються над даними, багато в чому залежить від використовуваних методів, обраних на попередньому етапі [5].

Третій етап - це власне застосування методів Data Mining. Сценарії цього застосування можуть бути найрізноманітнішими і можуть включати складну комбінацію різних методів, особливо якщо використовувані методи дозволяють проаналізувати дані з різних точок зору [5].

Наступний етап - перевірка побудованих моделей. Дуже простий і часто використовуваний спосіб полягає в тому, що всі наявні дані, які необхідно аналізувати, розбиваються на дві групи. Як правило, одна з них більшого розміру, інша - меншого. На більшій групі, застосовуючи ті чи інші методи Data Mining, отримують моделі, а на меншій - перевіряють їх. За різницею в точності між тестової та навчальної групами можна судити про адекватність побудованої моделі [5].

Останній етап - інтерпретація отриманих моделей людиною з метою їх використання для прийняття рішень, додавання одержаних правил і залежностей в бази знань і т. д. Цей етап часто включає використання методів, які перебувають на стику технології Data Mining і технології експертних систем. Від того, наскільки ефективним він буде, в значній мірі залежить успіх у вирішенні поставленої задачі [5].

1.4. Огляд інструментів SAS для текстової аналітики

Компанія SAS пропонує потужний набір інструментів, призначених для текстової аналітики, інтелектуального аналізу даних, прогнозного

моделювання. Перелік задач та методів текстової аналітики, що їх реалізують представлений на рисунку 1.2.

Рисунок 1.2 – Методи текстової аналітики і відповідні SAS інструменти

Методи текстової аналітики і відповідні SAS інструменти:

- SAS Crawler, SAS Search, Indexing використовується для збору або ідентифікації набору текстових матеріалів в Інтернеті, або в файловій системі, базі даних чи в іншому сховищу даних. Ці сирі дані можна використовувати в інших SAS інструментах текстової аналітики, наприклад в SAS Text Miner для подальшого аналізу. SAS Search і Indexing допоможе створити простий і складний індекс, щоб більш ефективно обробляти пошукові запити користувача, тобто отримати релевантну інформацію. Запити надсилаються к цим, завчасно створеним індексам, щоб повернути найбільш релевантні документи [10].
- SAS Enterprise Content Categorization використовується для представлення колекції документів у вигляді структурованої ієрархії категорій і підкатегорій, що має назву таксономія. Окрім категоризації документів,

цей інструмент може використовуватись для витягу фактів з них. Наприклад, статті з новинами можемо подати у вигляді завчасно визначеного набору категорій, це може бути політика, спорт, бізнес, фінанси і т. д. За допомогою цього інструмента досить легко витягти фактичну інформацію (події, місця, ім’я людей, дати, грошові суми і т. д.) [11].

- SAS Ontology Management використовується для інтеграції вже існуючих репозиторіїв документів на підприємстві і пошуку взаємозв’язків між ними. Цей інструмент може допомогти працівникам в області управління знаннями створити онтології і збудувати ієрархії семантичних взаємозв’язків між термінами для покращення пошуку і отримання інформації з репозиторіїв документів [10].
- SAS Text Miner використовується, щоб витягти ключові теми в текстових документах. Цей інструмент може згрупувати схожі документи на основі частоти термінів в каталозі документів, чи в окремому документі. Наприклад, текстові дані, отримані з контакт-центру, можна загрузити в SAS Text Miner, щоб провести їх автоматичну кластеризацію. Дані, що потрапили в один кластер, стосуються схожих проблем, що були описані клієнтами, а дані у різних кластерах стосуються різних проблем. Специфіку проблем можна зрозуміти, проаналізував терміни, що описують кожен кластер. Графічне представлення цих проблем і пов’язаних з ними термінів, подій і людей можна реалізувати за допомогою зв’язування концептів, яке продемонструє силу взаємозв’язку між подією і проблемою.
- SAS Text Miner дозволяє користувачу самостійно визначити теми. Документи можна оцінити за допомогою згадування визначених тем в них. При наявності цільової змінної за допомогою SAS Text Miner можливо побудувати класифікаційну модель з учителем або прогнозну модель. Результати прогнозної моделі, яка використовує кількісні вхідні

зміні, можливо покращити за допомогою тем, кластерів, чи правил, які витягуються з текстових коментарів за допомогою SAS Text Miner [9].

– SAS Sentiment Analysis Studio використовується для визначення настроїв до об’єкту в документ чи загального настрою до документа в цілому. Об’єктом може бути що завгодно, наприклад, продукт, атрибут продукту, бренд, людина, група чи навіть організація. Настрій може бути позитивним, негативним чи нейтральним (некласифікованим). Якщо немає пов’язаних з об’єктом термінів, що позначають настрій, настрій позначається як «некласифікований».

  Аналіз тональності тексту зазвичай застосовується до такого виду текстової інформації, як відгуки клієнтів о продуктах, брендах, організаціях, або опитування з причини тих, чи інших громадських подій (наприклад, президентські вибори). Такий вид інформації широко розповсюджений в соціальних мережах (Facebook, Twitter, YouTube) [11].

1.5 Висновки до розділу 1

Територіальні громади є складними системами, що характеризуються нестаціонарними і нелінійними процесами, тому досить складно знайти інструменти для їх автоматизованого управління. Тому необхідно створити систему підтримки прийняття рішень, яка була б у змозі опрацювати значні обсягів різнорідної та неструктурованої інформації, та забезпечувати зворотній зв'язок щодо ефективності управлінських рішень, які вже були прийняті.

Інтелектуальний аналіз даних дозволяє обробити великі обсяги неструктурованих і різнорідних еколого-економічних даних з різних джерел інформації, щоб представити їх у нормалізованому, структурованому вигляді, придатному для подальшого аналізу.

Компанія SAS пропонує потужний набір інструментів, призначених для текстової аналітики, інтелектуального аналізу даних, прогнозного моделювання: SAS Crawler, SAS Search, Indexing, SAS Enterprise Content Categorization, SAS Ontology Management, SAS Sentiment Analysis Studio, SAS Text Miner.

РОЗДІЛ 2
МЕТОДИКА ВИКОРИСТАННЯ ЗАСОБІВ ТЕКСТОВОЇ АНАЛІТИКИ ДЛЯ АНАЛІЗУ ТА КАТЕГОРИЗАЦІЇ ЕКОЛОГО-ЕКОНОМІЧНИХ ДАНИХ

2.1 Особливості збору та обробки еколого-економічних даних для цілей прогнозування розвитку територіальних громад

Соціально-економічний розвиток регіону - це категорія, яка відображає єдність і розвиток всіх регіональних аспектів, сфер і фаз суспільного відтворення на основі єдності продуктивних сил і соціально-економічних відносин, що з'являються в процесі життєдіяльності людини. Основні складові даної категорії визначаються економічною, соціальною, екологічною, демографічною сферами суспільного виробництва (рисунок 2.1) [12].

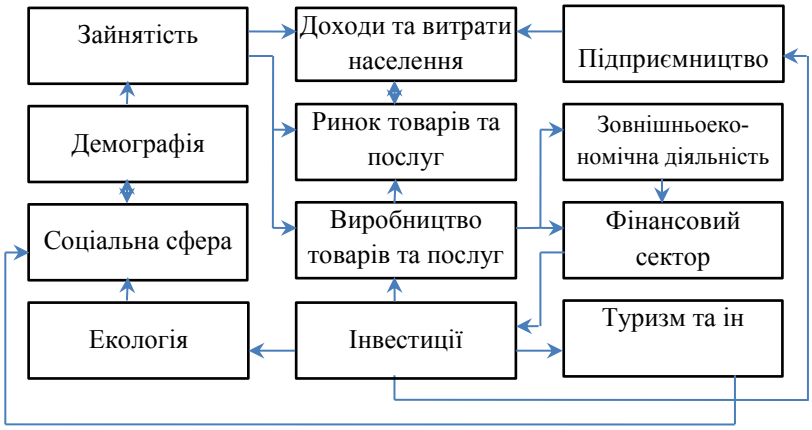

Рисунок 2.1 – Основні складові соціально-економічного розвитку регіону

Розвиток регіону як складної соціально-економічної системи характеризується величезною кількістю чинників, які можуть бути агреговані в наступні складові:

– економічна: інноваційний розвиток, інвестиційний клімат, економічний розвиток, інституційне забезпечення, розвиток фінансового сектору, виробництво товарів та послуг, зовнішньо-економічна діяльність;

– соціальна: економічна активність, науково-технічний потенціал, соціальний захист, соціальну рівновагу в суспільстві, екологічна безпека і гуманітарний розвиток [13].

На рисунку 2.2 приведена схема процесу прийняття рішень

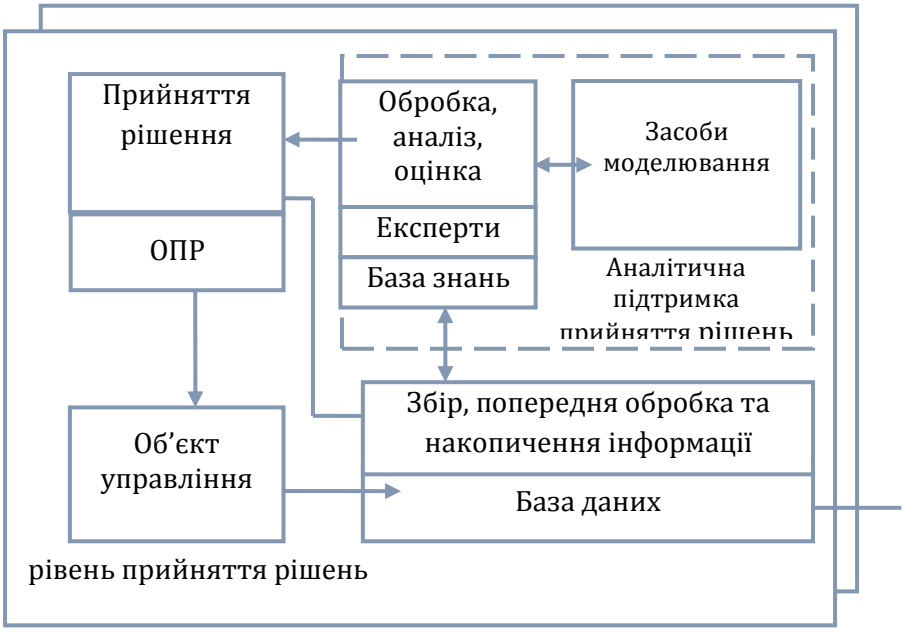

Рисунок 2. 2 – Основні елементи підтримки процесу прийняття рішень в управлінні соціально-економічними системами

На етапі збору, попередньої обробки та накопиченні інформації і виявленні характерних ознак об'єкта дослідження запропоновано автоматизоване опрацювання значних обсягів неструктурованої інформації за допомогою інструментів сімейства SAS.

Предметний домен S_0 «сільське господарство» має наступні складові S_{11} та S_{12}: <'культури', 'ринки'> згідно моделі опису складної ієрархічної системи. Рівні S_{1i} складаються у свою чергу з інших рівнів S_{2j}, наприклад: S_{11} = <'кукурудза', 'пшениця', 'ячмінь', 'зерно', 'гречка', 'соняшник', 'ріпак' >. Кожне джерело рівня S_{11} має над собою функціональні залежності, що перетворюють його кількісно чи якісно: $\phi_k(S_{1j}) \in$ *<'покупка', 'продаж', 'вирощування', 'селекція', 'транспортування', 'зберігання', 'переробка',*

‘використання’> та впливають на зовнішні параметри системи S_0 .

Синтез схеми у якій за відповідними шаблонами зв’язуються показники з економічного словника *(kpi)* та рівні ієрархії предметного домену. Для визначення трендів до застосованих вище показників додані типові показники з економічного тезаурусу (2.1):

$$kpi\_trends = kpi \times lvl \cap kpi \times dir \qquad (2.1)$$

де kpi=<arg_j>;

arg_j ∈ <споживання, вартість, дефіцит, попит, знижка, надлишок, інвестиції, вихід, ціна, квота, ризик, частка, ринкова вартість, субсидія, поставка, тариф, ставка податку, обсяг> та інші;

lvl=<arg_k>;

arg_k ∈<великий, низький>;

dir = <arg_l>;

arg_l ∈=<зріст, спад>.

Загальна модель вилучення фактів з текстів природною мовою: E=< всі текстові об'єкти (документи) у вхідних даних, всі правила, логічна функція (t_i, v_j), що приймає значення «Істина» якщо t_i задовольняє v_j >[14].

На рисунку 2.3 зображено ієрархію концепцій з галузі національної економіки.

Тип	H1	H2	H3	H4	Текст	Ключові слова
LI	A. Сільське господарство, лісове господарство та рибне господарство	01. Сільське господарство, мисливство та надання пов'язаних із ними послуг	01.1. Вирощу-вання однорічних і дворічних культур	01.11. Вирощування зернових культур (крім рису), бобових культур і насіння олійних культур	пшениця	пшениця
LI	A. Сільське господарство, лісове господарство та рибне господарство	01. Сільське господарство, мисливство та надання пов'язаних із ними послуг	01.1. Вирощування однорічних і дворічних культур	01.11. Вирощування зернових культур (крім рису), бобових культур і насіння олійних культур	кукурудза (на зерно)	зерно кукурудза

Рисунок 2.3 - Ієрархія концепцій галузі, сумісна з КВЕД, використана для опису та класифікації метаданих предметної області

2.2 Застосування засобів текстової аналітики для обробки еколого-економічних даних

Проблему інтелектуального аналізу тексту можливо описати в рамках тих кроків, які необхідно виконати практично в будь-якій задачі інтелектуального аналізу даних. Першим кроком є отримання ознак з колекцій документів, щоб можна було виконувати обчислення і застосовувати статистичні методології. Витягнуті ознаки повинні якимось чином відображати зміст документів. В ідеалі вони повинні фіксувати контент таким чином, щоб документи, які обговорюють схожі теми, але з іншою термінологією, мали аналогічні характеристики. Далі, необхідно сформулювати спосіб вимірювання відстаней між документами, де відстань будується, щоб вказати, як вони схожі за змістом. Як буде показано нижче, це не тільки дозволяє нам застосовувати методи класифікації і кластеризації, але також дозволяє формулювати стратегії для зменшення розмірності. З огляду на візуалізацію елементів в просторі з більш низькою розмірністю, ми можемо потім звернути нашу увагу на кластеризацию, дискримінантний аналіз [15].

Перша частина вилучення ознак - це попередня обробка словника або лексикону (сукупність усіх унікальних слів в колекції документів). Зазвичай це включає три частини: видалення стоп-слів, стемінг (визначення коренів слів) і визначення ваги термінів. До лексикону можна застосувати будь-яку, всі або жодну з цих трьох частин. Застосування і корисність цих методів - відкрите питання дослідження в області інтелектуального аналізу текстових даних, яке повинно представляти інтерес для статистичного співтовариства[15].

Стоп-слова - це загальні слова, що не додають значимого змісту документу. Ось деякі приклади: і, але, або. Стоп-слова можуть бути заздалегідь заданим списком слів або залежати від контексту документу.

Стемінг часто застосовується в області пошуку інформації, де метою є підвищення продуктивності системи та зменшення кількості унікальних слів.

Стемінг - це процес видалення суфіксів і префіксів, залишаючи корінь або основу слова[15]. Наприклад, слова «захищений», «захисти», «підзахисний» і «захист» будуть скорочені до кореневого слова «захист». Це має сенс, оскільки слова мають схоже значення. Однак в деяких випадках скорочене слово має геть інше значення. Скоріш за все, це не впливає на результати пошуку інформації, але може мати деякі небажані наслідки при класифікації і кластеризації. Стемінг і видалення стоп-слів зменшить розмір словника, тим самим заощадивши обчислювальні ресурси.

Один із способів кодування тексту - це підрахувати, скільки разів термін зустрічається в документі. Це називається частотним методом. Однак терміни з великою частотою не обов'язково більш важливі або володіють більш високою здатністю розпізнавання. Отже, ми могли б захотіти, щоб терміни мали вагові коефіцієнти по відношенню до локального контексту, документу або колекції документів.

Найпопулярнішим показником, мабуть, є IDF (обернена частота документа) - інверсія частоти, з якою слово зустрічається в документах колекції [16] (2.2).

$$IDF = \log(\frac{N}{df_i}) \tag{2.2}$$

де N - загальна кількість документів;

df_i - кількість документів, що містять термін і.

Існує розширення цього показника, який позначається як TF-IDF.Точне формулювання TF-IDF приведено нижче [16] (2.3).

$$w_{ij} = f_{ij} * \log(\frac{N}{df_i}) \tag{2.3}$$

де w_{ij} – ваговий коефіцієнт терміну і в документі j;

f_{ij} - кількість входжень терміну і в документі j;

N - загальна кількість документів;

df_i-кількість документів, що містять термін і.

З огляду на те, що лексикон був попередньо оброблений (чи ні), ми потім кодуємо вміст документів. Один з найпростіших способів зробити це - використати підхід з використанням векторного простору набору слів. Цей підхід має перевагу, тому що він не вимагає будь-якої обробки природної мови, такої як ідентифікація частини мови або мовний переклад. Колекція документів кодується як терм-документа матриця X з n рядками і p стовпцями, де n представляє кількість слів у лексиконі (повний або попередньо оброблений словниковий запас), а p - кількість документів. Таким чином, елемент x_{ij} матриці X містить кількість разів, скільки і-й термін зустрічається в j-ому документі (частота терміна). Як було написано вище, це також може бути зважена частота. Кодування колекції документів у вигляді матриці дозволяє використовувати можливості лінійної алгебри для аналізу колекції документів [17]. В таблиці 2.1 наведено приклад терм-документної матриці.

Таблиця 2.1 - Терм-документна матриця

	D1	D2	D3	D4	D5	D6	D7	D8	D9
He	2	2	0	1	2	1	1	2	1
Cash	1	0	0	0	0	1	0	0	0
Check	1	0	0	0	0	0	0	1	0
Bank	1	1	0	0	1	1	1	0	0
River	0	1	0	2	0	0	1	0	0
Boat	0	1	0	1	1	0	1	0	0
Be	0	1	0	0	1	0	0	0	0
On	0	1	1	0	0	0	0	0	0
Borrow	0	0	1	0	0	1	0	0	0
Credit	0	0	1	0	0	1	0	0	0
Floats	0	0	0	1	0	0	1	1	0
By	0	0	0	0	1	0	1	0	0
Dock	0	0	0	0	1	0	1	0	0
Near	0	0	0	0	1	0	1	0	0
Parade	0	0	0	0	0	0	0	1	1
Route	0	0	0	0	0	0	0	1	1

В таблиці 2.1. позначені документи:

D1: deposit the cash and check in the bank;

D2: the river boat is on the bank;

D3: borrow based on credit;

D4: river boat floats up the river;

D5: boat is by the dock near the bank;

D6: with credit, I can borrow cash from the bank;

D7: boat floats by dock near the river bank;

D8: check the parade route to see the floats;

D9: along the parade route.

Підхід векторного простору було розширено за рахунок включення порядку слів в сенсі пар або трійок слів. Замість однієї матриці кодується кожен документ як матрицю, що називається матрицею близькості, так що тепер у нас є р матриць для роботи. По-перше, нам потрібно провести додаткову попередню обробку документів. Вся пунктуація видаляється (наприклад, коми, крапки з комою, двокрапки, тире і т. д.), а всі розділові знаки в кінці речення перетворюються в крапку. Крапка вважається словом в лексиконі. Таким чином, кожна матриця близькості має n рядків і n стовпців [17].

В таблиці 2.2 показаний приклад матриці близькості для документа, що складається з одного речення «Створення сучасних технологій прийняття рішень.».

Таблиця 2.2 - Матриця близькості

	створення	сучасних	технологій	прийняття	рішень	.
створення	0	1	0	0	0	0
сучасних	0	0	1	0	0	0
технологій	0	0	0	1	0	0
прийняття	0	0	0	0	1	0
рішень	0	0	0	0	0	1
.	0	0	0	0	0	0

Для вирішення завдань інтелектуального аналізу текстових даних, пов'язаних з кластеризацією, класифікацією і пошуком інформації, необхідно застосувати поняття відстані або подібності між документами. Найбільш використовуваною мірою при інтелектуальному аналізі текстових даних і пошуку інформації є косинус кута між векторами, що представляють документи [18]. Припустимо, у нас є два вектори документа $\vec{a}$ і $\vec{q}$, тоді косинус кута між ними, θ, визначається як (2.4):

$$\cos(\theta) = \frac{\vec{a}^T \vec{q}}{||\vec{a}||_2 \, ||\vec{q}||_2} \tag{2.4}$$

де $||\vec{a}||_2 = \sqrt{\sum_{i=1}^{n} a_i^2}$ - L_2 норма вектору $\vec{a}$.

Зверніть увагу, що великі значення цього показника вказують на близьке розташування документів, а менші значення вказують на те, що документи знаходяться далі один від одного.

Міра косинуса - це міра подібності, а не відстань. Зазвичай нам зручніше працювати з відстанями, але ми можемо легко перетворити міру подібності в відстані. По-перше, припустимо, що ми організували наші подібності в позитивно визначену матрицю C, де ij-й елемент цієї матриці вказує на схожість i-го і j-го документів. Тоді один із способів перетворити це значення в Евклідову відстань - використовувати наступну формулу [19] (2.5):

$$d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}} \tag{2.5}$$

Зверніть увагу: якщо два документа співпадають ($c_{ii} = c_{jj}$), то відстань між ними дорівнюватиме нулю.

Простір, в якому знаходяться документи, зазвичай має досить велику розмірність. З огляду на набір документів, поряд з відповідною матрицею

відстаней, часто буває цікаво знайти зручний простір меншої розмірності для виконання подальшого аналізу. Це може бути вибрано для полегшення візуалізації, кластеризації та класифікації. Можна сподіватися, що, застосовуючи зменшення розмірності, можна видалити шум з даних і краще застосовувати методи статистичного аналізу даних для виявлення взаємозв'язків, які можуть існувати між документами [20].

Спочатку розглянемо особливо цікаву проекцію (спосіб зменшити кількість вимірювань), яку можна розрахувати безпосередньо з матриці термін-документ. Можна використовувати відому теорему лінійної алгебри, щоб отримати набір корисних проекцій за допомогою розкладання за сингулярними числами. Це стало відомо в інтелектуальному аналізі текстових даних та їх обробки природної мови як LSA (латентно-семантичний аналіз).

Нехай X позначає матрицю дійсних чисел розміром $m \times n$ і ранг дорівнює r, де $m \geq n$ і, отже, $r \leq n$.

Розкладання по сингулярним числах дозволяє нам записати матрицю як добуток трьох матриць [20] (2.6):

$$X = U \times S \times V^T \qquad (2.6)$$

де U - матриця розміру $m \times n$;

S - діагональна матриця $n \times n$;

V^T - матриця розміру $n \times n$.

Стовпці U називаються лівими сингулярними векторами, $\{u_k\}$, і утворюють ортонормований базис, так що $u_i \cdot u_j = 1$ для $i = j$ і $u_i \cdot u_j = 0$ в іншому випадку. Рядки V^T містять елементи правих сингулярних векторів $\{v_k\}$. Елементи S відмінні від нуля тільки на діагоналі і називаються сингулярними значеннями. Таким чином, $S = \text{diag}(s_1, ..., s_n)$. Крім того, $s_k > 0$ для $1 \leq k \leq r$ і $s_k = 0$ для $(r + 1) \leq k \leq n$. Зазвичай, порядок сингулярних векторів визначається сортуванням сингулярних значень по спадаючій з найвищим сингулярним значенням у верхньому лівому індексі S-матриці [20].

Одним з важливих результатів SVD для матриці X є те, що (2.7)

$$X^{(l)} = \sum_{k=1}^{l} u_k \, s_k v_k^T \tag{2.7}$$

- найближча до X матриця рангу l. Термін «найближча» означає, що $X^{(l)}$ мінімізує суму квадратів різниці елементів X і $X^{(l)}$ (2.8)

$$\sum_{ij} \left| x_{ij} - x_{ij}^{(l)} \right| \to min \tag{2.8}$$

Один із способів розрахунку SVD - це спочатку обчислити V^T і S шляхом діагоналізації X^TX (2.9):

$$X^TX=VS^2V^T \tag{2.9}$$

Потім обчислити U наступним чином (2.10):

$$U=XVS^{-1} \tag{2.10}$$

де (r + 1), ..., n стовпців V, для яких $s_k = 0$, ігноруються при матричному множенні. Вибір решти n-r сингулярних векторів в V або U може бути обчислений з використанням процесу ортогоналізації Грама-Шмідта.

Між PCA (метод головних компонент) і SVD існує прямий зв'язок у тому випадку, коли головні компоненти обчислюються з коваріаційної матриці. Якщо від кожного стовпчика матриці X відняти середнє значення цього стовпчика, то $X^TX = \Sigma_i g_i g_i^T$ пропорційно коваріаційній матриці змінних g_i [20].

Відповідно до рівняння 2.8 діагоналізація X^TX дає V^T, яка також дає головні компоненти $\{g_i\}$. Отже, праві власні вектори $\{v_k\}$ збігаються з головними компонентами $\{g_i\}$. Власні значення X^TX еквівалентні s_k^2, які пропорційні дисперсії головних компонентів.

Застосування SVD в аналізі даних має схожість з аналізом Фур'є. Як і у випадку з SVD, аналіз Фур'є включає розширення вихідних даних по ортогональному базису[20] (2.11):

$$x_{ij} = \sum_k c_{ik}\, e^{i2\pi jk/m} \tag{2.11}$$

Зв'язок з SVD можна явно проілюструвати, нормалізував вектор $\{e^{i2\pi jk/m}\}$ і назвавши його v'_k (2.12):

$$x_{ij} = \sum_k b_{ik}\, v'_{jk} = \sum_k u'_{ik} s'_k v'_{jk} \tag{2.12}$$

яке породжує матричне рівняння (2.13):

$$X=U`S`V`^T \tag{2.13}$$

Це матричне рівняння аналогічно рівнянню 2.6. На відміну від SVD, однак, навіть якщо $\{v'_k\}$ є ортонормованим базисом, $\{u'_k\}$ в цілому не ортогональні.

Сингулярний розклад застосовується по відношенню до частотної матриці, що дозволяє спроектувати документи в простір меншої розмірності. SVD розмірність, що згенерована, за своїм змістом найкраще відображає підпростір термінів за критерієм найменших квадратів.

На рисунку 2.4 приведена двомірна діаграма розсіювання документів:

D1: deposit the cash and check in the bank;

D2: the river boat is on the bank;

D3: borrow based on credit;

D4: river boat floats up the river;

D5: boat is by the dock near the bank;

D6: with credit, I can borrow cash from the bank;

D7: boat floats by dock near the river bank;

D8: check the parade route to see the floats;

D9: along the parade route.

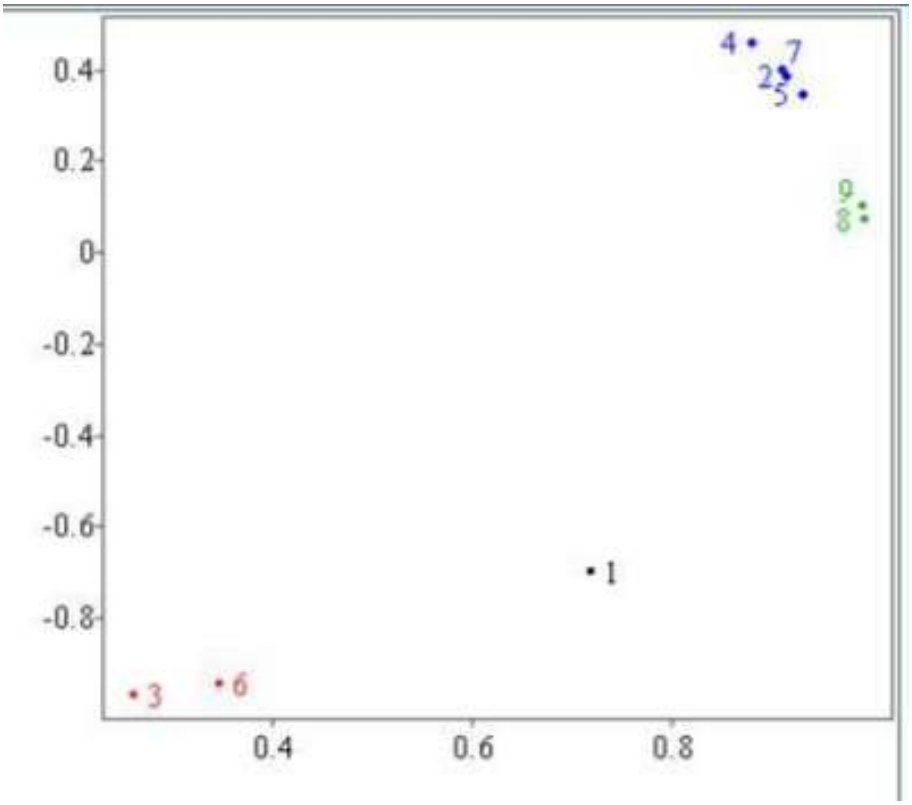

Рисунок 2.4 - Двовимірна діаграма розсіювання документів

Документ 1 ближче до документа 3, ніж до документа 2. З іншого боку, документ 5 безпосередньо пов'язаний з документами 2, 4, 7 – це означає, що проекція намагається розмістити схожі за змістом документи поруч, не дивлячись на те, що у них не так багато спільних термінів. SVD розкладання використовує в даному випадку двомірний підпростір замість шістнадцяти мірного (один вимір доводиться на один термін).

Латентно-семантичний аналіз (LSA) - це алгебро-статистичний метод, який визначає значення слів та схожість речень, використовуючи інформацію про використання слів у контексті. Він зберігає інформацію про те, які слова використовуються у реченнях, зберігаючи при цьому інформацію про спільні слова серед інших речень. Велика кількість спільних слів між реченнями

позначає, що ці речення більш семантично пов'язані. Метод LSA може одночасно відображати значення слів та значення речень. Він усереднює значення слів, що міститься в реченнях, щоб дізнатися значення цього речення. Метод LSA використовує розклад за сингулярними значенням (SVD) для виявлення семантично схожих слів та речень [21].

LSA має три основні обмеження. Перше обмеження в тому, що він використовує лише інформацію у вхідному тексті та не використовує інформацію про світові знання. Друге обмеження в тому, що не використовує інформацію про порядок слів, синтаксичні відносини або морфології. Така інформація використовується для визначення значень слів та текстів. Третє обмеження в тому, що продуктивність алгоритму знижується з більшими и неоднорідними даними. Зниження продуктивності спостерігається, оскільки SVD, який є досить важким алгоритмом, використовується для виявлення подібності [21].

Усі методи, основані на LSA, використовують три основні кроки. Ці кроки зображені на рисунку 2.5.

Рисунок 2.5 - Алгоритм латентно-семантичного аналізу

2.3 Категоризація еколого-економічних даних для використання їх під час побудови причинно-наслідкових та когнітивних моделей

Для виділення категорій в еколого-економічних даних використано методами кластеризації. Вирізняють декілька груп вказаних методів. Ієрархічні методи: ці методи створюють кластери, рекурсивно розділяючи екземпляри або зверху вниз, або знизу догори. Серед ієрархічних метдів вирізняють такі методи як агломераційна ієрархічна кластеризація та розділова ієрархічна кластеризація. За агломераційної ієрархічної кластеризації - кожен об'єкт спочатку представляє власний кластер. Потім кластери послідовно об'єднуються, до тих пір, поки не буде отримана бажана кластерна структура. Розділова ієрархічна кластеризація - всі об'єкти спочатку належать одному кластеру. Потім кластер розділюється на підкластери, які послідовно розділюються на свої власні підкластери. Цей процес продовжується до тих пір, доки не буде отримана бажана кластерна структура [23].

Результатом використання ієрархічних методів є дендрограма (рисунок 2.6).

Рисунок 2.6 – Приклад дендограми

Дендограма - це деревоподібна структура, в яку записують послідовність зливань та розділень, в якій вертикальна лінія представляє відстань між кластерами. Відстань між вертикальними лініями та відстань між кластерами прямо пропорційна, тобто чим більша відстань, тим більше кластерів, ймовірно, буде неоднакових. Кластеризація об'єктів даних отримується через відсічення дендрограми на бажаному рівні подібності.

Злиття або розділення кластерів виконується відповідно до деякої міри подібності, вибраної для оптимізації деякого критерію (наприклад, суми квадратів).

Методи ієрархічної кластеризації можуть бути додатково розділені відповідно до способу розрахунку міри подібності:

Метод одиночного зв'язку («метод найближчого сусіда») - метод, який рахує відстань між двома кластерами, як найкоротшу відстань від будь-якого члена одного кластера до будь-якого члена іншого кластера [22] (2.14):

$$\min \{d(a, b): a \in A, b \in B\} \quad (2.14)$$

де d(a, b) - відстань між елементами a і b, що належать кластерам A і B.
Метод одиночного зв'язку має недолік, відомий як «ефект ланцюжка»: кілька точок, які утворюють міст між двома кластерами призводять до об'єднання цих двох кластерів в один кластер.

Метод повного зв'язку («метод далекого сусіда») - метод, який рахує відстань між двома кластерами, як найдовшу відстань від будь-якого члена одного кластера до будь-якого члена іншого кластера [22] (2.15):

$$\max \{d(a, b): a \in A, b \in B\} \quad (2.15)$$

Метод середнього зв'язку («методом мінімальної дисперсії») – метод, який вважає, що відстань між двома кластерами дорівнює середній відстані від

будь-якого члена одного кластера до будь-якого члена іншого кластера [22] (2.16):

$$\frac{1}{|A||B|}\sum_{a\,\epsilon\,A}\sum_{b\,\epsilon\,B} d(a,b) \qquad (2.15)$$

де d(a, b) - відстань між елементами a і b, що належать кластерам A і B;

|A| - кількість елементів кластера A.

Недоліком методу середнього зв’язку може бути поділ подовжених кластерів і злиття частин сусідніх подовжених кластерів.

Центроїдний метод – метод, який вважає, що відстань між кластерами дорівнює відстані між центроїдами цих класів [22] (2.17):

$$|| c_A - c_B|| \qquad (2.17)$$

де c_A і c_B – центроїди A і B.

Метод Уорда. На відміну від інших методів кластерного аналізу, для оцінки відстаней між кластерами тут використовуються методи дисперсійного аналізу. У якості відстані між кластерами береться приріст суми квадратів відстаней об'єктів до центру кластера, одержуваного в результаті їх об'єднання [22] (2.18):

$$\Delta=\sum_i (x_i - \bar{x})^2 - \sum_{x_i\,\epsilon\,A}(x_i - \bar{a})^2 - \sum_{x_i\,\epsilon\,B}(x_i - \bar{b})^2 \qquad (2.18)$$

На кожному кроці алгоритму об'єднуються такі два кластери, які призводять до мінімального збільшення дисперсії. Цей метод застосовується для задач з близько розташованими кластерами.

Методи секціонування: ці методи переміщують екземпляри, переміщаючи їх з одного кластера в інший, починаючи з початкового поділу. Такі методи зазвичай вимагають, щоб кількість кластерів було попередньо

встановлено користувачем. Для досягнення глобальної оптимальності кластеризації методом секціонування потрібен процес перебору всіх можливих секцій. Оскільки це неможливо, деякі жадібні евристики використовуються в формі ітеративної оптимізації [23].

До методів секціонування належать алгоритми мінімізації помилок. Ці алгоритми, які, як правило, добре працюють з ізольованими і компактними кластерами, є найбільш інтуїтивно зрозумілими і часто використовуваними методами. Основна ідея полягає в тому, щоб знайти структуру кластеризації, яка мінімізує певний критерій помилки, який вимірює «відстань» кожного елемента кластера до його репрезентативного значення. Найвідоміший критерій - це сума квадратів помилок (SSE), що дорівнює сумі квадратів різниці між екземплярами кластеру і їх репрезентативними значеннями. Сума квадратів помилок може бути глобально оптимізована шляхом перебору всіх можливих секцій, що забирає дуже багато часу, або шляхом надання приблизного рішення (не обов'язково веде до глобального мінімуму) з використанням евристики. Останній варіант - найбільш поширена альтернатива [23].

Найпростішим і найбільш часто використовуваним алгоритмом, що використовує критерій квадрата помилки, є алгоритм K-середніх. Цей алгоритм розбиває дані на K кластерів (C_1, C_2, ..., C_K), представлених їх центрами. Центр кожного кластера обчислюється як середнє значення всіх елементів, що належать цьому кластеру. Алгоритм починається з початкового набору центрів кластерів, обраних випадковим чином. На кожній ітерації кожному елементу присвоюється найближчий до нього центр кластера відповідно до евклідової відстані між ними. Потім центри кластерів перераховуються знову. Центр кожного кластера обчислюється як середнє значення всіх елементів, що належать цьому кластеру (2.19):

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \qquad (2.19)$$

де N_k - кількість елементів, що належать кластеру k;

μ_k - середнє значення кластера k.

Існує декілька умов збіжності. Наприклад, пошук може припинитися, якщо сума квадратів помилок не зменшиться після переміщення центрів. Це вказує на те, що даний розділ є локально оптимальним. Також можуть використовуватися інші критерії зупинки, такі як перевищення заздалегідь заданої кількості ітерацій.

Інший алгоритм кластеризації, який намагається мінімізувати SSE, - це K-medoids. Цей алгоритм дуже схожий на алгоритм K-середніх, але він відрізняється від тим що кожен кластер представлений найбільш центрованим об'єктом в кластері, а не неявним середнім значенням, яке може не належати кластеру [23].

Серед неієрархічних алгоритмів, які не ґрунтуються на відстані, слід виділити EM-алгоритм. У ньому замість центрів кластерів передбачається наявність функції щільності ймовірності для кожного кластеру з відповідним значенням математичного очікування і дисперсією. Перед початком алгоритму висувається гіпотеза про вигляд розподілів, які оцінити в загальній сукупності даних складно. EM алгоритм є основним методом пошуку оцінки максимальної правдоподібності параметра, що лежить в основі розподілів з множини заданих даних. Вважається, що всі змінні є незалежними, і всі дані мають k спільних розподілів. Основний алгоритм розділений на два кроки [24].

E-алгоритм (крок очікування) [24] (2.20):

$$z_{ij} = \frac{p(x_j|y_i)p(y_i)}{p(x_j)} \tag{2.20}$$

де $p(x_j|y_i)$ визначає ймовірність виникнення x_j в кластері y_i;

$p(y_i)$ визначає ймовірність виникнення y_i;

$p(x_j)$ визначає ймовірність виникнення x_j.

M-алгоритм (крок максимізації) [24] (2.21-2.22):

$$u_i = \frac{\sum_j^n z_{ij} x_i}{\sum_j^n z_{ij}} \quad (2.21)$$

$$\sigma_i^2 = \frac{\sum_j^n z_{ij} (x_j - u_i)^2}{\sum_j^n z_{ij}} \quad (2.22)$$

де u_i – середнє розподілу i;

σ_i^2 – дисперсія розподілу i;

z_{ij} - розрахована ймовірність спостереження j, що належить кластеру i.

Якщо оцінене зверху значення правдоподібності менше зазначеного порогового значення або кількість ітерацій дорівнює максимальному числу, то робота алгоритму припиняється і отримаємо остаточну кластеризацію.

Значення правдоподібності для даного алгоритму виражається функцією (2.23):

$$L = \sum_j^n log \sum_i^k p(x_j|c_i) p(c_i) \quad (2.23)$$

Метод К-середніх використовується для виділення груп об'єктів в економіці, при аналізі даних, а також в інформаційно-пошукових системах.

Метод ієрархічної кластеризації використовується при зборі статистичних даних і реалізований в статистичних пакетах. Так само використовується при кластеризації текстових документів.

EM-алгоритм застосовується в інформаційно-пошукових системах для кластеризації великого обсягу даних [24].

2.4 Висновки до розділу 2

У другому розділі описано основні кроки в задачі інтелектуального аналізу даних. Попередня обробка словника - це визначення значущих термінів. Зазвичай це включає три частини: видалення стоп-слів, стемінг (визначення

коренів слів) і визначення ваги термінів. На наступному кроці колекція документів кодується як терм-документа матриця і формулюється спосіб вимірювання відстаней між документами, де відстань будується, щоб вказати, як вони схожі за змістом. У розділі описано метод сингулярного розкладу матриці для зменшення розмірності (дозволяє зменшити шум у даних) та латентно-семантичного аналізу і методи кластеризації (ієрархічні методи, K-середніх, EM-алгоритм).

РОЗДІЛ 3

ОПИС РОЗРОБЛЕНОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

3.1. Опис структури системи

В результаті виконання магістерського дослідження розроблено систему аналізу та категоризації еколого-економічних даних для прогнозування розвитку територіальних громад . Структуру системи зображено на рисунку 3.1.

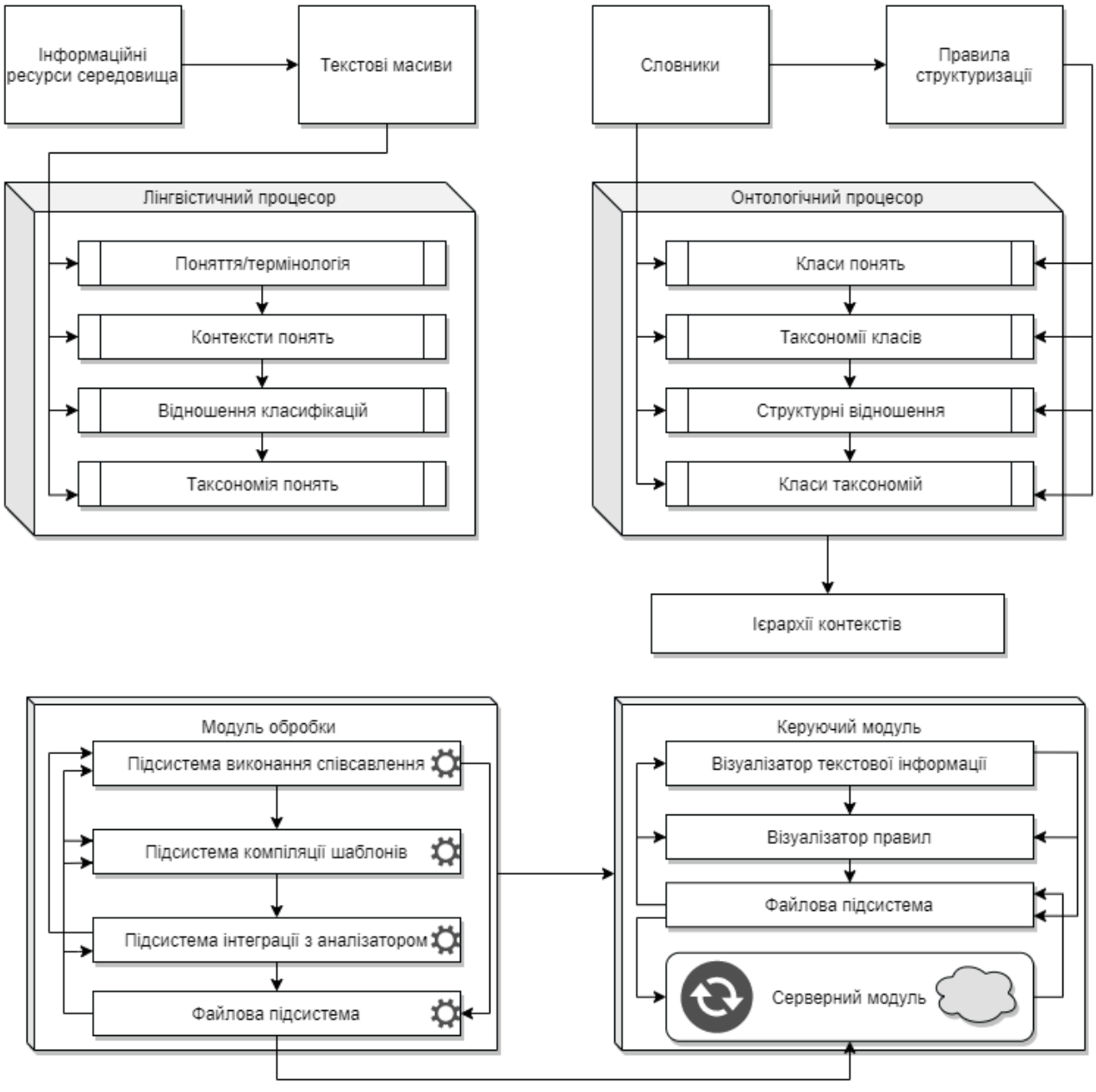

Рисунок. 3.1 - Загальна архітектура розробленої системи

За допомогою даної системи можна видобути, категоризувати та проаналізувати інформацію з різних наборів документів. Також є можливість використати дану систему і для інших текстових даних, якщо налаштувати її основні та допоміжні компоненти.

Документи, що необхідно обробити, зберігаються у внутрішній базі даних, з якої користувач може використовувати їх для проведення певних дій або аналізу. Якщо необхідно проаналізувати дані або виконати їх категоризацію користувач має завантажити їх у спеціальний додаток в середовищі SAS, що дозволить підвищити ефективність обробки шляхом зберігання їх у пам’яті [25].

Завантаживши дані до основної системи, користувач одержує основні характеристики текстової колекції: кардинальне число множин, розподіл документів за об’ємом та розширенням та ін.

Так як для побудови математичних моделі не можливо використвувати неструктуровані дані, необхідно виконати роцедури попередньої обробки текстової інформації: об’єднання синонімів, стеммінг, відокремлення частин мови та перетворення тексту у чисельні вектори (зважження), та видалення певних термів, що не впливають на зміст текстів [25].

Після того, як попередня обробка даних закінчилась, користувач має визначити основні терміни, що мають вплив на аналіз, якщо їх не вистачає у створеному списку понять. Проводиться аналіз тональності текстів, будуються зв’язки між основними поняттями, з яких отримуються базові тематики, які користувач може відрегулювати щоб одержати кращий результат. Категорії створюються з тематик, що найчастіше зустрічались в документах за відповідними булевими правилами [25].

Користувач може отримати результати:

- у вигляді SAS коду, що можна редагувати;
- у вигляді графіку розподілу документів по категоріям;
- у вигляді таблиці в якій документи розподілені по відповідним категоріям [25].

3.2. Основні функції розробленого програмного забезпечення

На рисунку 3.2 зображено SAS Enterprise Miner, в якому можна відокремити набір даних та необхідні бібліотеки пам’яті CAS (Cloud Analytical Services). Перед тим, як провести початковий аналіз даних, необхідно завантажити їх у CAS-пам’ять, для чого, ці дані потрібно перетворити у формат HDFS (Hadoop Distributed File System) для їх завантаження в пам’ять CAS.

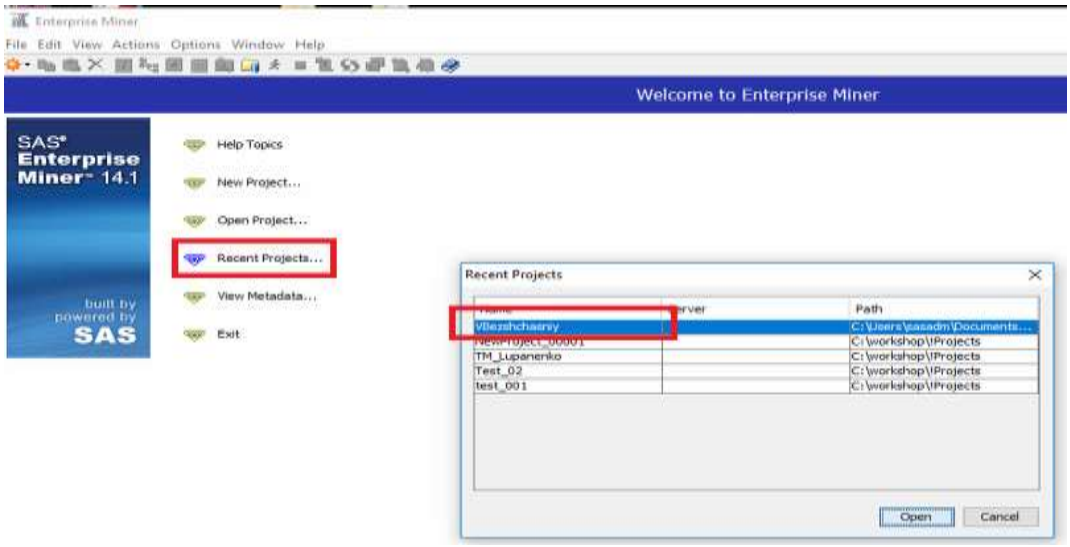

Рисунок 3.2 – Початкове вікно SAS Enterprise Miner

На рисунку 3.3 зображено вікно вибору типу діаграми, яку потрібно обрати.

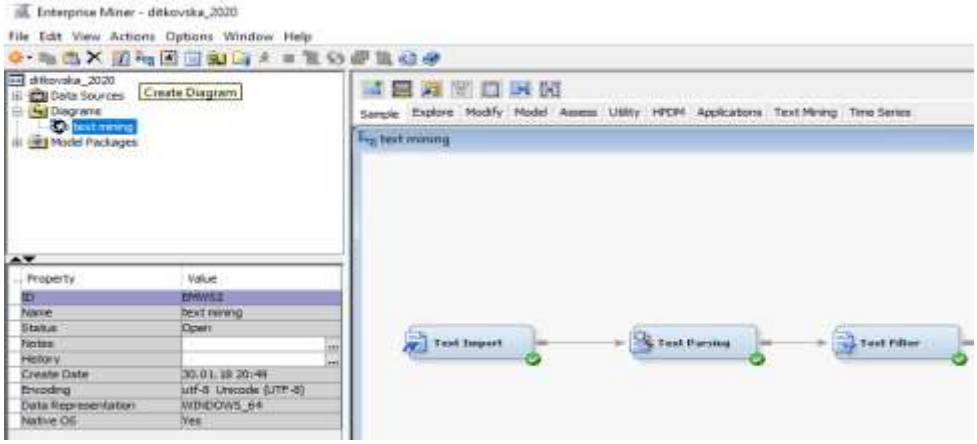

Рисунок 3.3 – Діаграма послідовності виконання проекту

На рисунку 3.4 побудована діаграма ієрархічної і EM-кластеризації, виконаних в системі SAS Enterprise Miner.

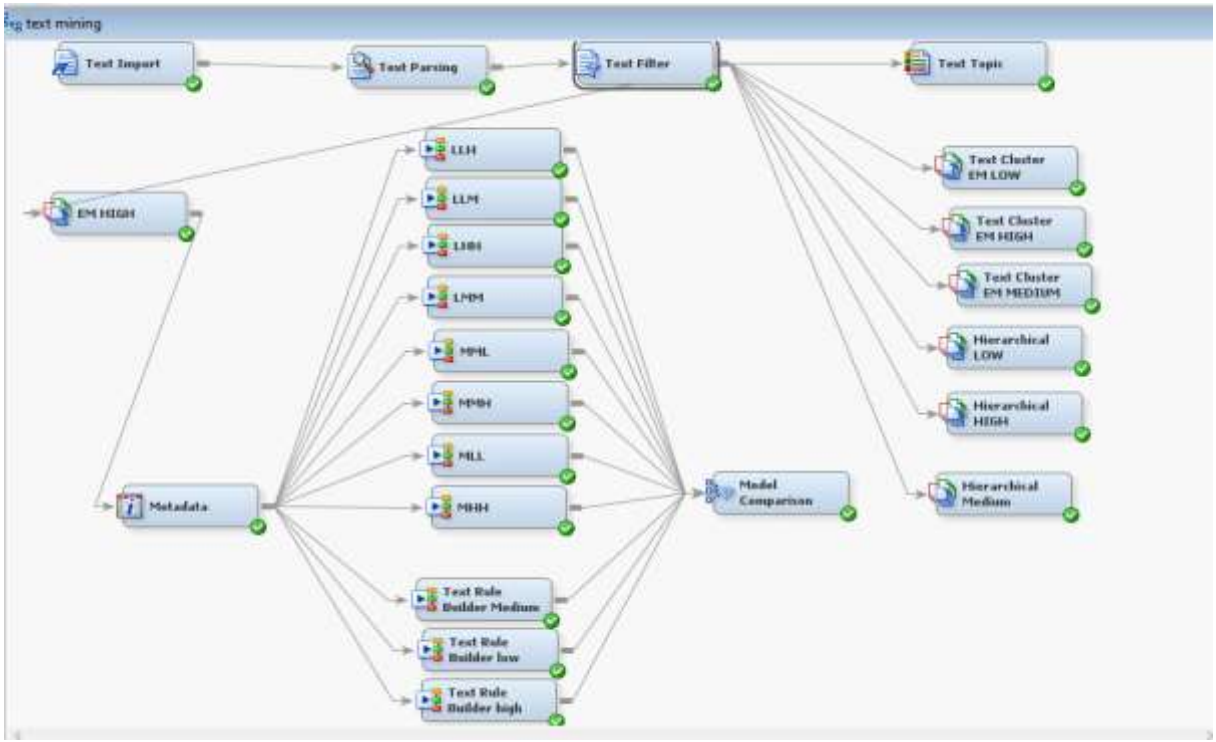

Рисунок 3.4 - Діаграма кластеризації в системі SAS Enterprise Miner

Розмістивши файли в папці, можна виконати запуск програми, як зображено на рисунку 3.5.

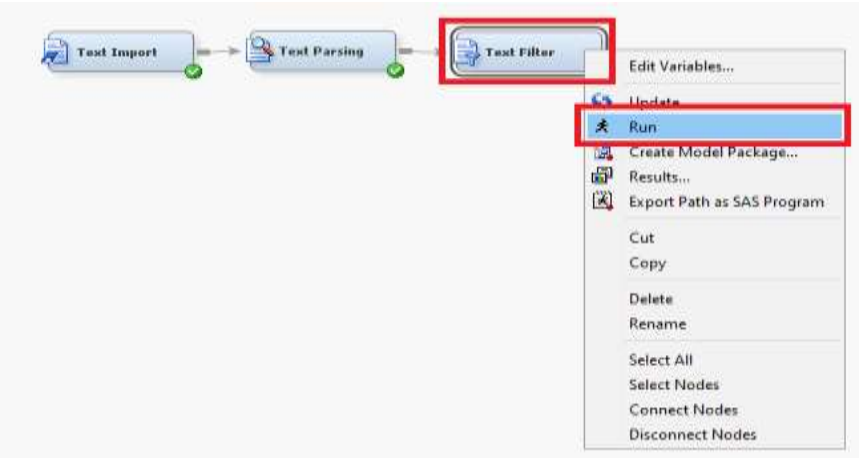

Рисунок 3.5 – Запуск фільтрації термінів

Якщо потрібно додати нові файли для обробки, їх необхідно розмістити в тій самій папці, що й основні, та виконати перезапуск програми таким самим способом, як показано на рисунку 3.5. Після виконання програми, користувачу будуть доступні результати обробки тексту у вигляді таблиці з метаданими, а також гістограми, які показують розподіл документів відповідно до їх розширення.

Після того, як фільтрація завершилась, обираємо ключове слово, та запускаємо програму. В результаті виконання отримаємо візуалізацію у вигляді концептуальних зв'язків досліджуваного з іншими поняттями (приклад зображено на рисунку 3.6). [25]

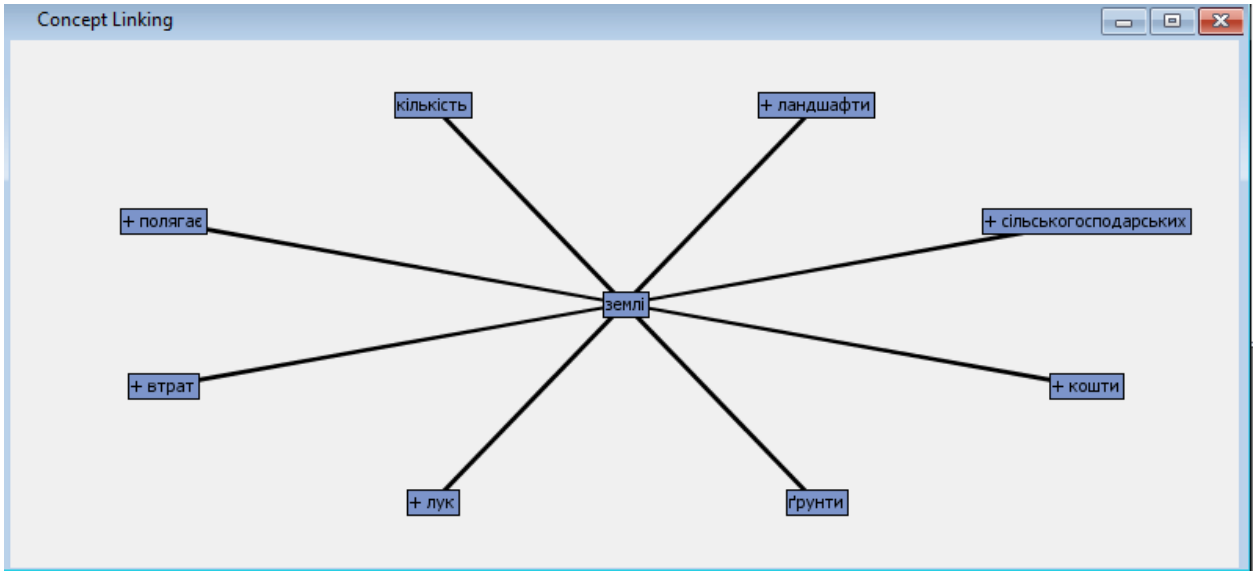

Рисунок 3.6 –Концептуальні зв'язки терміна «земля» з екологічними факторами

Кожне поняття, пов'язане з ключовим, можна розкрити, проаналізувавши наступні концептуальні зв'язки, пов'язані вже з цим поняттям. Це дуже важливо коли розглядаються рішення щодо еколого-економічного обгрунтування стратегічних планів сталого розвитку територіальних громад. Оскільки це дозволить робочій групі обгрунтувати вибір визначити сильні та слабкі сторони, можливості та загрози розвитку під час проведення загального SWOT- аналізу спираючись на результати дослідження думки мешканців, яку вони висловлюють, зокрема в соціальних мережах та на інформаційних ресурсах місцевого самоврядування.

Приклад наведено на рисунку 3.7, на якому розкриті наступні поняття після «землі». Зокрема, «земля» розглядається не лише з точки зору природного ресурсу (його стану та екології), а й як джерело для розвитку соціально-економічної системи територіальної громади (ресурс, який має вартість та може бути використаний).

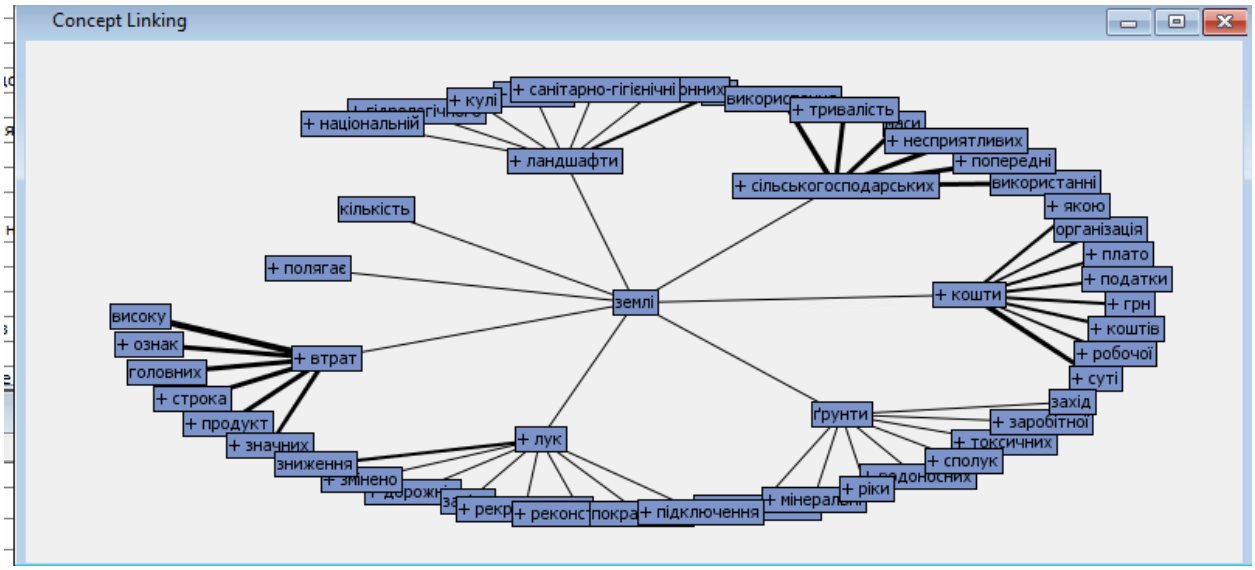

Рисунок 3.7 – Розширено концептуальні зв’язки терміну «земля»

Для завантаження повного словнику колекції текстів, використовується команда обрати “Result” меню фільтрації (рисунок 3.8).

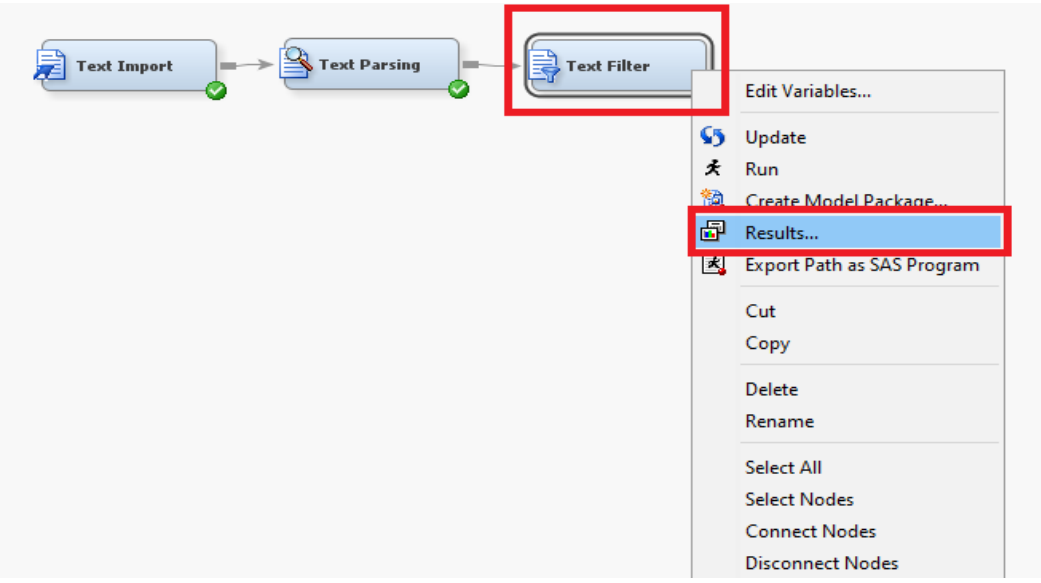

Рисунок 3.8 – Отримання словника термінів для колекції текстів

В представленій колекції більше 50 понять, що з'являються у пошуку більше 60 разів. Список основних понять наведено у таблиці 3.1 та в додатку

Таблиця 3.1 - Корпус слів, з частотою появи більше 60

Term	Role	Attr	Status	Weight	Freq	# Docs
+ україни	Prop	Alpha	Keep	0.34	6919.0	131.0
+ стан	Noun	Alpha	Keep	0.39	2335.0	106.0
+ час	Noun	Alpha	Keep	0.34	1182.0	100.0
+ розвитку	Noun	Alpha	Keep	0.41	3249.0	89.0
+ використання	Noun	Alpha	Keep	0.35	1651.0	86.0
+ система	Noun	Alpha	Keep	0.36	1448.0	81.0
+ території	Noun	Alpha	Keep	0.37	1069.0	80.0
+ рівні	Noun	Alpha	Keep	0.31	600.0	79.0
+ мають	Noun	Alpha	Keep	0.36	742.0	79.0
+ довкілля	Noun	Alpha	Keep	0.32	1149.0	77.0
+ виробництва	Noun	Alpha	Keep	0.40	1076.0	76.0
+ проблема	Noun	Alpha	Keep	0.40	1031.0	74.0
системи	Noun	Alpha	Keep	0.36	1673.0	74.0
+ природних	Noun	Alpha	Keep	0.38	1150.0	73.0
+ ресурсів	Noun	Alpha	Keep	0.37	1307.0	73.0
Рівень	Noun	Alpha	Keep	0.38	683.0	72.0
+ країни	Noun	Alpha	Keep	0.40	479.0	71.0
+ господарства	Noun	Alpha	Keep	0.33	560.0	70.0
+ можуть	Noun	Alpha	Keep	0.41	572.0	70.0
+ видів	Noun	Alpha	Keep	0.36	809.0	70.0
+ охорони	Noun	Alpha	Keep	0.38	1299.0	70.0
+ заходів	Noun	Alpha	Keep	0.37	1177.0	70.0
+ управління	Noun	Alpha	Keep	0.35	1388.0	69.0
+ рівня	Noun	Alpha	Keep	0.39	872.0	68.0
+ впливу	Noun	Alpha	Keep	0.38	817.0	68.0
+ вода	Noun	Alpha	Keep	0.39	1227.0	68.0
+ зміни	Noun	Alpha	Keep	0.36	755.0	67.0
Можна	Noun	Alpha	Keep	0.43	785.0	67.0
+ середовища	Noun	Alpha	Keep	0.42	1595.0	67.0
Також	Prop	Alpha	Keep	0.29	208.0	67.0
+ рік	Noun	Alpha	Keep	0.38	753.0	67.0
+ вонь	Noun	Alpha	Keep	0.41	565.0	67.0
+ природный	Adj	Alpha	Keep	0.39	1351.0	67.0
+ контроль	Noun	Alpha	Keep	0.34	838.0	66.0
+ створення	Noun	Alpha	Keep	0.38	927.0	66.0
+ основних	Noun	Alpha	Keep	0.37	527.0	66.0
+ діяльності	Noun	Alpha	Keep	0.40	1350.0	66.0
+ повітря	Noun	Alpha	Keep	0.33	1084.0	65.0
+ водить	Verb	Alpha	Keep	0.40	1078.0	65.0
+ захисту	Noun	Alpha	Keep	0.37	629.0	65.0
+ відповідно	Noun	Alpha	Keep	0.36	741.0	65.0
+ забезпечення	Noun	Alpha	Keep	0.38	1306.0	65.0
роботи	Noun	Alpha	Keep	0.34	486.0	65.0

Проаналізувавши утворений корпус слів, можна зробити висновок, про те що найбільше турбує мешканців територіальної громади: стан навколишнього середовища, який погіршується під впливом забруднення, але треба розвивати виробництво, яке забезпечить роботою, однак, плануючи розвиток регіону, треба забезпечити відтворення природного середовища за одночасного забезпечення економічного зростання. Тобто, використання такого підходу може бути застосоване в процесі обгрунтування стратегічного вибору громади.

3.3 Аналіз результатів роботи програми

Для перевірки працезданості системи було використано 177 різноманітних інформаційних джерел, розміщених в мережі Інтернет: від статей до аналітичних висновків. В результаті отримано категоризацію повідомлень та дописів мешканців, що характеризують важливість урахування екологічних факторів для розвитку громади та усвідомлення їх впливу на економіку. На рисунку 3.9 зображено схему аналізу інформації.

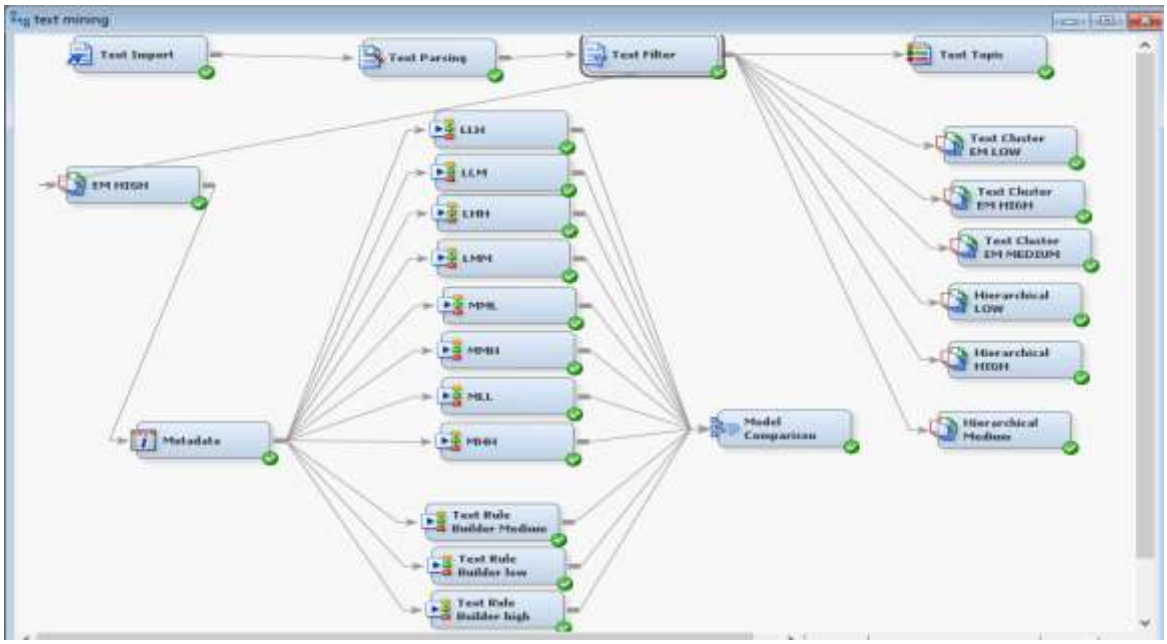

Рисунок 3.9 – Схема процесу аналізу даних

Основним кроком у текстовій аналітиці є збір, відокремлення та формування словника термінів. В процесі виконується стеммінг, парсинг, видалення зайвих слів, визначення частин мови та контроль правописання. Також виділяються поняття, що були задані на минулому етапі [25].

На наступному кроці колекція документів кодується як терм-документа матриця і формулюється спосіб вимірювання відстаней між документами, де відстань будується, щоб вказати, як вони схожі за змістом [25].

На рисунку 3.10 зображено концептуальні зв’язки терміна “території”.

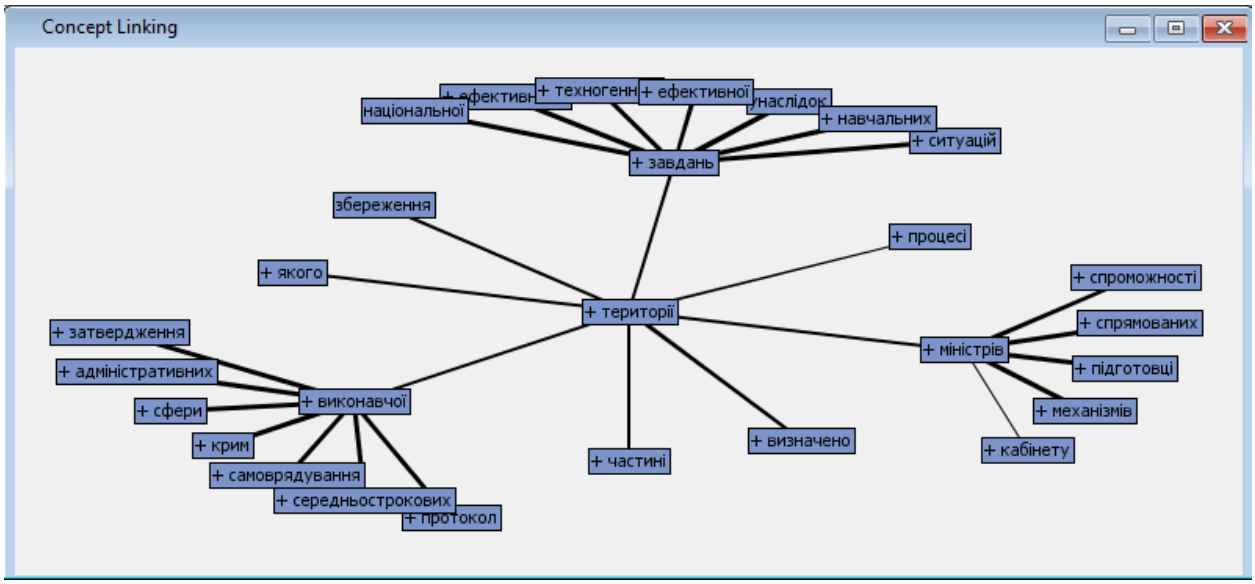

Рисунок 3.10 – Концептуальні зв’язки терміна “території”

Як видно з рисунку 3.10, на якому зображено карта взаємозв’язків для кожного терміну, з терміном “території” найсильніший зв’язок з терміном “виконавчої” [25]. Це пов’язано, скоріш за все, з тим, що саме на виконавчу гілку влади населення покладає виконання основних обов’язків з забезпечення сталого розвитку територій. Виконавча влада поряд із місцевим самоврядування сьогодні є основними ланками забезпечення реалізації державної політики та ефективності реформи децентралізації в регіонах.

Як видно з отриманих результатів, термін “території” спостерігається у 80 документах зі 177 та має частоту появи 1069 разів, що показує кількість разів, що дане слово з’являється в усіх текстах.

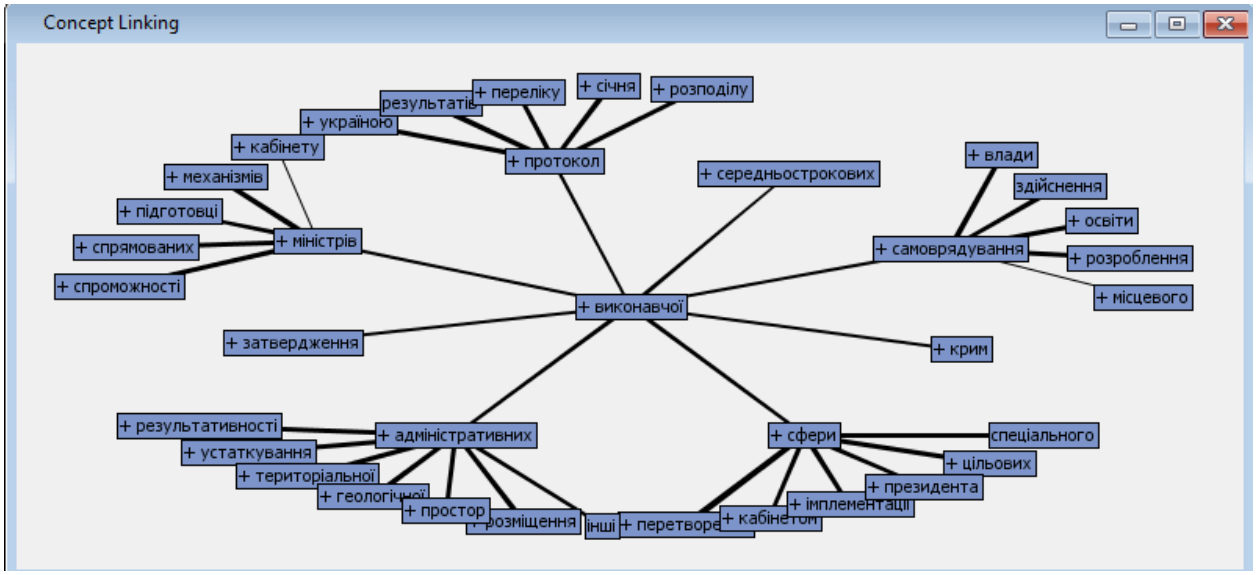

Рисунок 3.11 – Концептуальні зв’язки терміна “виконавча”

Слово “виконавча”, що сильніше пов’язане з терміном “території”, зустрічається у 53 текстах з 177, та має частоту появи 697 разів.

Маючи таку статистику, можна сказати, що близько 30% так чи інакше стосувались виконавчих органів місцевого самоврядування.

Термін “економіка” з’являється у 51 документі з 177 з частотою появи 894 разів. На рисунку 3.12 наведено візуалізацію концептуальних зв’язків даного терміна.

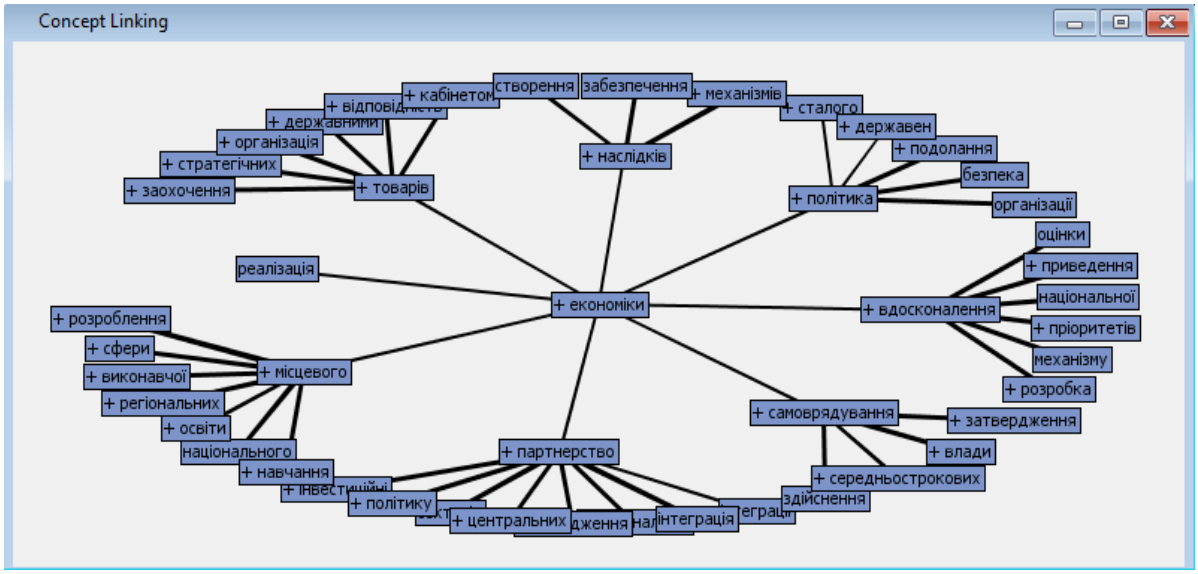

Рисунок 3.12 – Концептуальні зв’язки терміна “економіка”

Як видно з рисунку 3.12, поширеною думкою серед населення є та, що саме економіка є джерелом забезпечення сталого розвитку, потребує стратегічних реформ та активізації участі місцевого самоврядування у визначенні пріоритетних напрямків економіки громад.

Термін політика (рисунок 3.13) з'являється у 57 документах. А показник частоти появи дорівнює 1219.

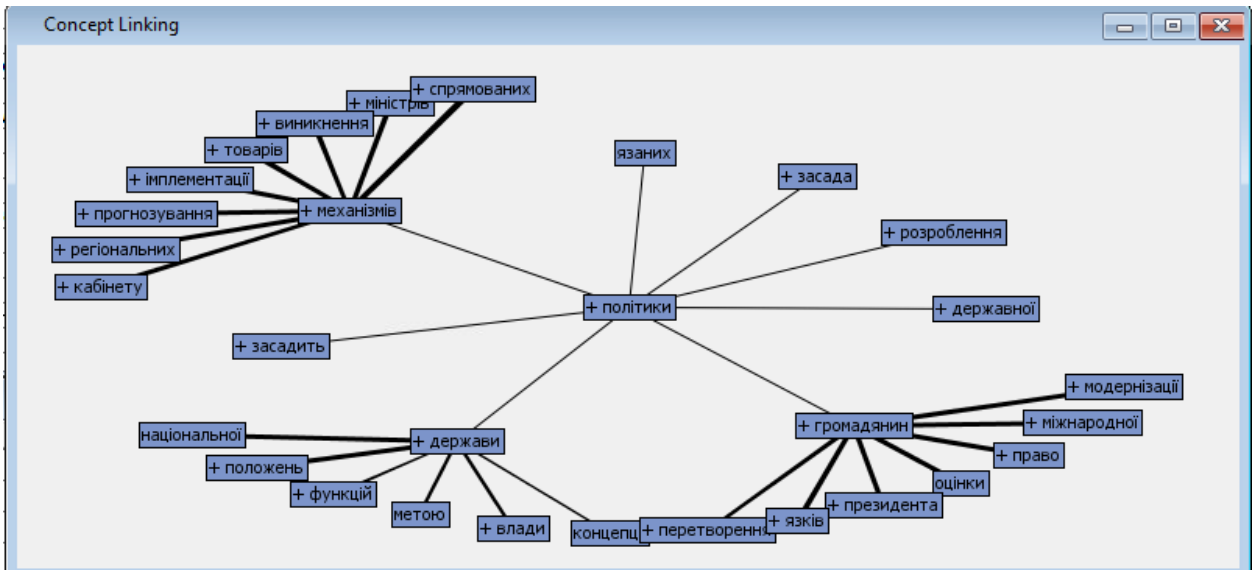

Рисунок 3.13 – Концептуальні зв'язки терміна "політика"

Термін екологія (рисунок 3.14) з'являється у 45 документах. А показник частоти появи дорівнює 432.

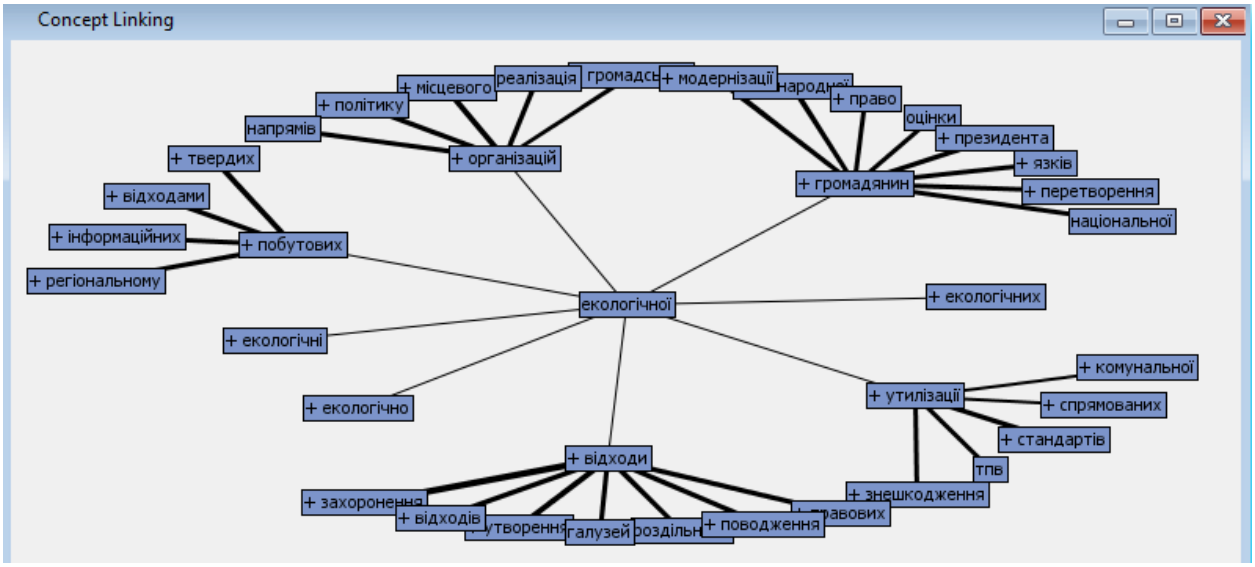

Рисунок 3.14 – Концептуальні зв'язки терміна "екологія"

Аналізуючи екологічну складову розвитку територіальних громад слід відзначити, що населення не байдуже до проблеми переробки відходів та сміття. Термін відходи (рисунок 3.15) зустрічається у 52 із 177 документів і загальна кількість появи поняття відходи в цілому наборі документів дорівнює 1240 разів.

Тому, можна зробити висновок, що з екологічних проблем, найбільш важливими вважаються проблеми відходів, їх утилізації чи переробки.

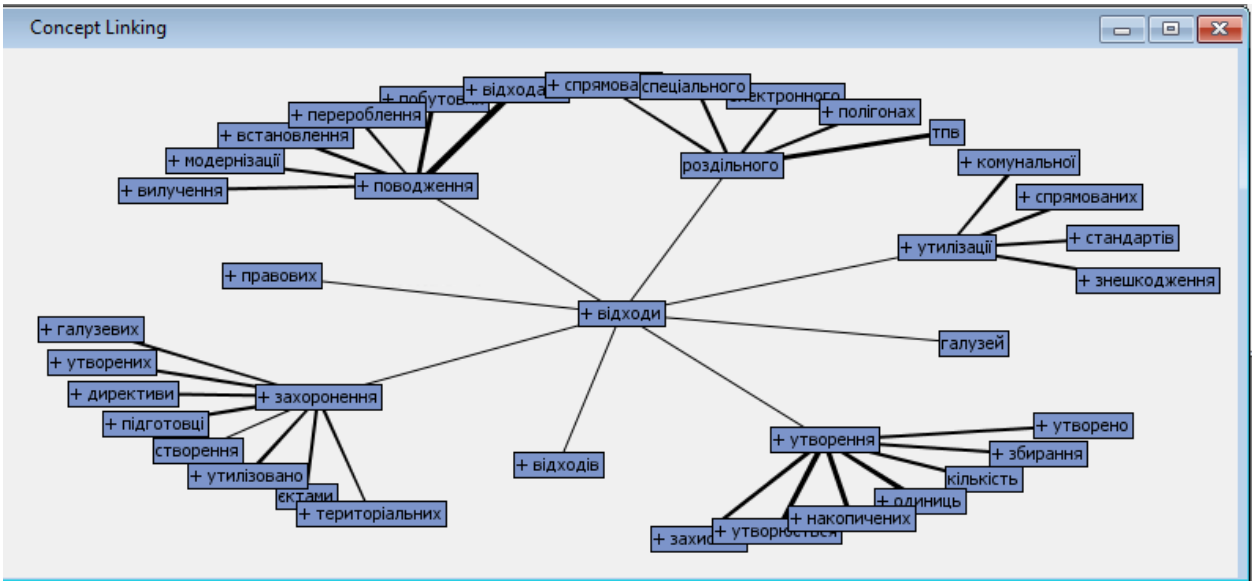

Рисунок 3.15– Концептуальні зв’язки терміна “відходи”

Для кластеризації колекції текстів використовується ієрархічний метод, що має реалізацію в SAS Enterprise Miner. Ієрархічні методи створюють кластери, рекурсивно розділяючи екземпляри або зверху вниз, або знизу догори. Алгоритм заснований на використанні власних сингулярних чисел (SVD – singular value decomposition), що трансформуються у вагові коефіцієнти з частотної матриці термів корпусу документів (таблиця 3.2)[26].

Результат методу ієрархічної кластеризації представляється у вигляді дерева, в якому батьківські вузли містять як мінімум два вузли наслідника, в свою чергу наслідники-вузли також мають свої піддерева [25].

Ієрархічна кластеризація використовує метод Уорда з мінімізацією

дисперсію, за цим методом відстань між двома кластерами обчислюється за формулою (3.1)[25]:

$$D = \frac{(u_1 - u_2)^{|}(u_1 - u_2)}{\frac{1}{n_1} + \frac{1}{n_2}} \quad (3.1)$$

де u_1 та u_2 математичні сподівання кластерів;

n_1 та n_2 розміри кластерів.

Було отримано 6 кластерів, які описано в таблиці 3.2.

Таблиця 3.2 - Ієрархія кластерів у вигляді таблиці з описом кожного з термів в режимі High SVD resolution

Номер кластеру	Опис кожного з термів	Кількість документів	Частка в колекції документів
1.0	території +господарства її +або їх +видів вони чи +охорони тому +україни +є +україні +зокрема +довкілля	46	0,26
2.0	реалізація порядку державна с. Основними м. національної р. Показники результати державного період проведення збереження стану	36	0,20
3.0	+із проблеми +відходи +ніж вже для року ще +екології	33	0,19
4.0	+млн т р +вчора землі+нагадаємо т. +тис +ринку листопада +компанії	27	0,15
5.0	потенціал результати період умови проведення системи рівень якщо після при +його +може їх +розвитку лише	19	0,11
6.0	м. національної рівень системи т. Проблеми при цьому +мр станом +даними +ринку	16	0,09

На рисунку 3.16 зображено відстань між кластерами у двовимірному просторі.

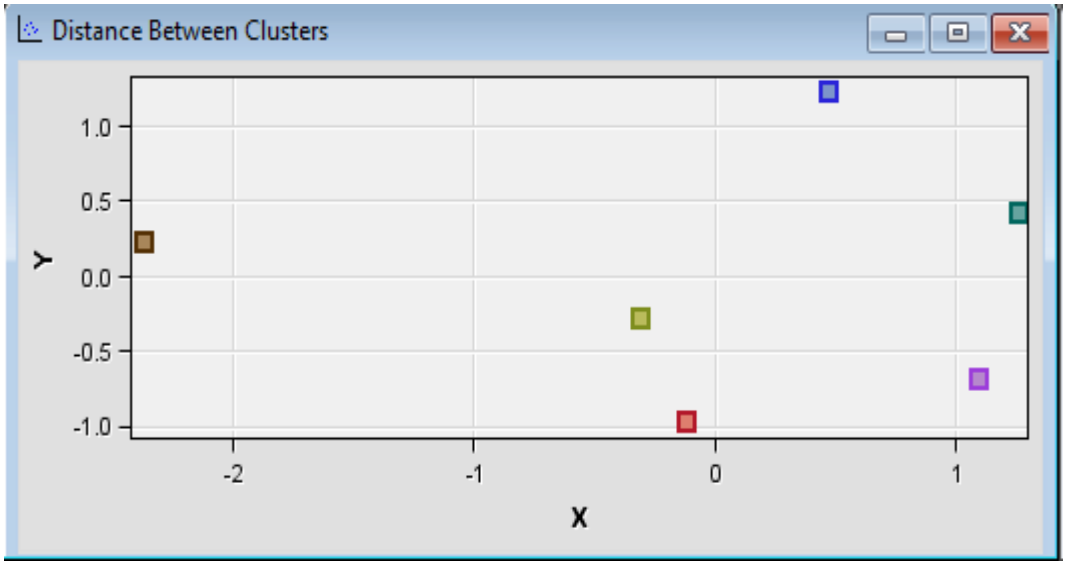

Рисунок 3.16 – Відстань між кластерами у двовимірному просторі

Є можливість вибору трьох варіантів ієрархічної кластеризації:

- При низькому (low SVD resolution), але швидкому варіанті кластеризації зазвичай достатньо розміру SVD простору, що пояснює приблизно ⅔ варіабельності матриці термів (рисунок 3.17) [25];
- При середньому (medium SVD resolution) варіанті кластеризації зазвичай достатньо розміру SVD простору, що пояснює приблизно ¾ варіабельності матриці термів (рисунок 3.18) [25];
- При найкращому (high SVD resolution) варіанті кластеризації зазвичай достатньо розміру SVD простору, що пояснює приблизно ⅚ варіабельності матриці термів (рисунок 3.19) [25].

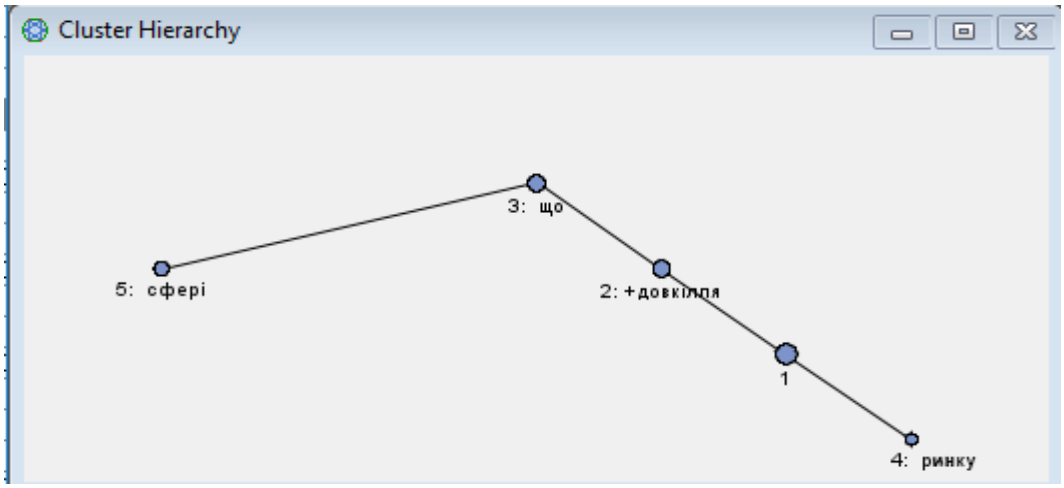

Рисунок 3.17 - Ієрархія кластерів у вигляді графу (low SVD resolution)

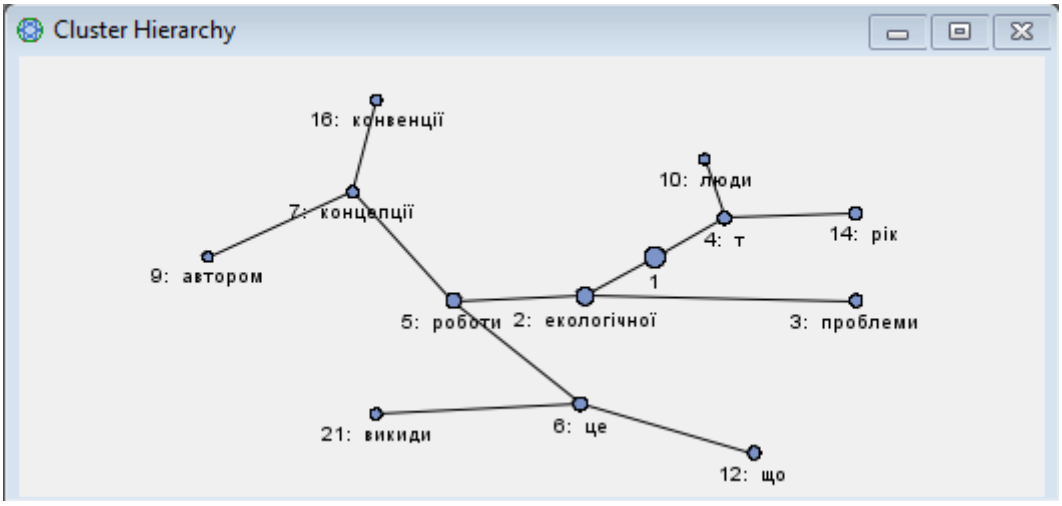

Рисунок 3.18 – Ієрархія кластерів у вигляді графу (medium SVD resolution)

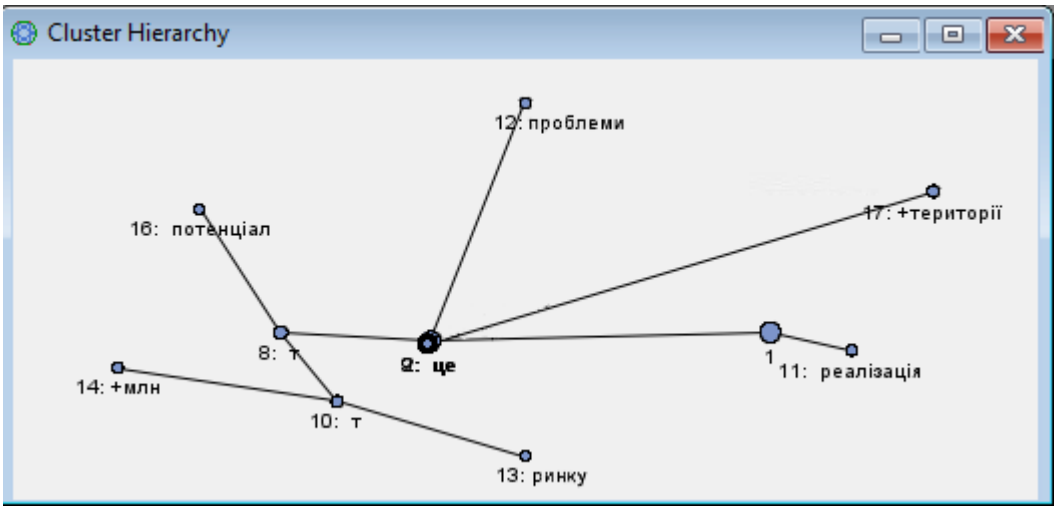

Рисунок 3.19 - Ієрархія кластерів у вигляді графу (high SVD resolution)

Порівняння роботи трьох режимів ієрархічної кластеризації описано в таблиці 3.3.

Таблиця 3.3 - Порівняння режимів ієрархічної кластеризації

	Режим ієрархічної кластеризації	Розмір SVD простору	Кількість кластерів	Час обчислення (в секундах)
1	Low SVD resolution	5	3	35
2	Medium SVD resolution	17	7	40
3	High SVD resolution	100	6	51

Після отримання кластерів, кожен наступний новий документ ми можемо віднести до існуючих кластерів за допомогою елементу Text Rule Builder. За допомогою цього елементу створюються набір правил, згідно яких прогнозується цільова змінна. Цей компонент має додаткові налаштування.

Generalization Error – визначає якість правил, з метою уникнення проблеми перенавчання [25].

Можливі значення: Very Low, Low, Medium (default), High, Very High.

Для запобігання перенавчання ставлять більше значення для складності, наприклад, Very High, але в цьому випадку можливі ситуації, коли дійсно корисні правила не будуть знайдені[25].

Purity of Rules – задає значення p-value, за допомогою якого визначається включати терм до правила чи ні.

Можливі значення: Very Low (p<.17), Low (p<.05), Medium (default, p<.005), High (p<.0005),Very High (p<.00005).

Exhaustiveness – визначає кількість згенерованих правил, чим більша кількість правил, тим більше необхідно часу на обчислення.

Можливі значення: Very Low, Low, Medium (default), High, Very High.

Визначаємо найкраще налаштування за допомогою метрики misclassification rate (частка неправильно класифікованих значень) (3.2).

$$\text{Misclassification rate} = \frac{false\ positives + false\ negatives}{total\ instances} \qquad (3.2)$$

В таблиці 3.4 наведено результати моделювання для 27 варіантів налаштувань.

Перевагами застосування кластерного аналізу у розробленій системі є те, що такий підхід дозволяє краще зрозуміти досліджувану предметну область, спростити подальшу обробку даних та їх використання у процесі прийняття рішень, оскільки до кожного визначеного кластеру можна застосувати окремий метод аналізу, прийнятний саме для об’єктів, що входять до нього. Крім того, за допомогою кластерного аналізу дослідник може виявити нетипові об’єкти, які радикально відрізняються від тих, що потрапили до кластерів, що дає

можливість відзначити їх позитивні та негативні особливості, що також дуже важливо при дослідженні розвитку еколого-економічних систем.

Таблиця 3.4 - Результати моделювання для 27 варіантів налаштувань

Generalization Error	Purity of Rules	Exhaustiveness	Misclassification rate
Very Low	Very Low	Very High	**0.05084745762711865**
Very Low	Very Low	Medium	0.062146892655367235
Very Low	Medium	Very High	0.0847457627118644
Very Low	Medium	Medium	0.096045197740113
Very Low	Very Low	Very Low	0.1016949152542373
Very Low	Medium	Very Low	0.12994350282485875
Medium	Very Low	Very High	0.13559322033898305
Very Low	Very High	Medium	0.15254237288135594
Medium	Medium	Very High	0.15254237288135594
Medium	Medium	Medium	0.15819209039548024
Very Low	Very High	Very High	0.15819209039548024
Medium	Very Low	Medium	0.1638418079096045
Very Low	Very High	Very Low	0.1638418079096045
Medium	Very Low	Very Low	0.1751412429378531
Medium	Medium	Very Low	0.1751412429378531
Very High	Very Low	Very High	0.2033898305084746
Medium	Very High	Very High	0.20903954802259886
Very High	Medium	Very High	0.21468926553672316
Medium	Very High	Medium	0.22033898305084745
Medium	Very High	Very Low	0.22598870056497175
Very High	Medium	Medium	0.23728813559322035
Very High	Very Low	Medium	0.23728813559322035
Very High	Very High	Medium	0.24293785310734464
Very High	Very High	Very High	0.24293785310734464
Very High	Very Low	Very Low	0.2655367231638418
Very High	Very High	Very Low	0.2655367231638418
Very High	Medium	Very Low	0.2655367231638418

Найменше значення Misclassification rate = 0.05 спостерігається, коли Generalization Error = Very Low, Purity of Rules= Very Low, Exhaustiveness= Very High.

При цих налаштуваннях було отримано наступні правила для категорізації.

cluster =1

(OR, (AND, (OR, "концентрації" , "концентрації" , "концентраційну"), (NOT, "чи")))

cluster=5

(OR, (AND, (OR, "сша2" , "сша"), (OR, "інституцій" , "інституцій" , "інституції" , "інституції")), (AND, (OR, "theo" , "them" , "then" , "they" , "the" , "they" , "thes"), "рівень"), (AND, (NOT, (OR, "умов" , "ум" , "умов" , "умо" , "умову")), (OR, "a." , "a")))

cluster=4

(OR, (AND, "т" , (NOT, (OR, "є." , "є"))), (AND, (OR, "comp" , "come" , "come" , "com", "comf"), (NOT, (OR, "є." , "є"))), (AND, (NOT, (OR, "віді" , "віду" , "відє" , "відь" , "відо" , "від")), (OR, "компанії" , "компаніїї" , "копані" , "компанії" , "компані" , "компані")))

cluster =3

(OR, (AND, "пізніх" , (OR, "ізз" , "із")), (AND, "викиди" , "радою"), (AND, (OR, "підгалузей" , "підгалузі")), (AND, "механізму" , (OR, "кліматично" , "кліматично" , "кліматичного" , "кліматичною" , "кліматичної" , "кліматичної"), "сфері"), (AND, (OR, "витриму" , "витримує")), (AND, (OR, "екологічність" , "екологічніст" , "екологічність" , "екологічній" , "екологічній"), (OR, "проживанню" , "проживання" , "промивання" , "проживанням")), (AND, "механізму" , (OR, "екологогічні" , "екологічні")))

cluster=2

(OR, (AND, (NOT, (OR, "повітрю" , "повітр" , "повітря" , "повіт" , "повітро")), (OR, "річок" , "річок")), (AND, (OR, "лісів" , "лісів" , "лісі"), (NOT, (OR, "політиками" , "політики" , "політиками" , "політики5"))), (AND, (OR, "on" , "on"), (NOT, (OR, "приз" , "прид" , "прин" , "пре" , "прид" , "преб" , "переть" , "прем" , "преш" , "пред" , "прев" , "прут" , "прип" , "приа" , "при" , "прем"))), (AND, (OR, "уп" , "упп")))

cluster =6

(OR, (AND, (OR, "рівень" , "системи")))

Аналізуючи отримані результати можна зробити висновки про те, що в кластер найбільшої активності включає ключові слова, пов’язані з проблемами екології та механізмами створення комфортних мов для життя, можна зробити висновок про те, що увагу при формуванні Стратегічного плану розвитку такої громади слід приділити більше уваги пошуку джерел фінансування заходів з утилізації відходів, впровадження контролю за раціональним використанням земельних ресурсів.

Отже, в використання система аналізу та категоризації еколого-економічних даних для прогнозування розвитку територіальних громад дозволяє на основі аналізу цитованості окремих термінів та словосполучень визначити пріоритети мешканців територіальних громад, дослідити їх зміну щодо забезпечення екологічної складової економічних реформ, формування відповідних інвестиційних стратегій. Впровадження пропонованої методики дозволить не лише обгрунтовувати формування матриць SWOT-аналізу, а й карту «Ланцюжок цінностей», яка дозволить поглиблювати дослідження еколого-еколонічної ситуації в різних територіальних громадах, виявляти точки зростання, зміну структури економіки та її вплив на довкілля, потенційних лідерів та успішні проекти розвитку.

3.4 Висновки до розділу 3

В цьому розділі розглянуто прикладну систему аналізу еколого-економічних даних з використанням SAS Enterprise Miner.

В результаті було отримано систему, що може категоризувати неструктуровані еколого-економічні дані, було отримано правила для категорізації. Проаналізувавши 177 різних джерел інформації, можна сказати, що значна кількість новин, пов’язаних з екологією, стосується відходів, їх переробки та утилізації. Це можна пояснити погіршенням загальної екологічної

ситуації в країні.

Новини економіки значною мірою пов’язані із сільськогосподарськими землями і врожайністю, що свідчить про інтерес до ситуації на аграрному ринку та необхідність раціонального використання земельних ресурсів, збереження їх родючості.

Значні сподування на щодо покращення ситуації населення покладає на органи місцевого самоврядування та успішність децентралізації.

Однією з переваг даної системи є можливість налаштувати автоматичний пошук в наявних даних, їх розподіл на різні категорії. Після обробки інформації користувач отримує не тільки результати категоризації, а також аналіз зв’язків між ключовими термінами.

По результатам роботи системи, можна зробити висновок, що ця система може знайти застосування у різноманітних напрямах екології або економіки, для приватних та державних інститутів, що займаються напрямом аналізу своїх сфер. Економія часу на пошук та висвітлення основних термів є значною, що дозволяє зменшити загальний час на прийняття управлінських та стратегічних рішень.

Використаний програмний продукт SAS Text Miner є хмарним рішенням, що дозволяє використовувати дану систему на майже будь-якому комп’ютері, маючи мінімальний набір можливостей. Це також дозволяє проводити велику кількість складних обчислень, виконувати аналіз значних обсягів даних незалежно від потужності наявного комп’ютеру.

РОЗДІЛ 4
РОЗРОБКА СТАРТАП-ПРОЕКТУ

В останні роки набув великої популярності такий вид малого підприємництва як стартап.

Стартап-проект – це комерційний проект, що найчастіше знаходиться в стані активної розробки, або нещодавно вийшов на ринок. Між стартапом та малим бізнесом є спільні риси, але вони все ж таки відрізняються: стартапу більш характерні інноваційні риси, такі як активне використання нових технологій та алгоритмів розробки, оригінальність – та, насправді, не завжди доцільність – ідеї та активний ріст на початку проекту. Проте, неможливо відхилити те, що стартапи стали одним з кроків розвитку сучасного бізнесу. При всьому цьому проект може бути як масштабного характеру, так і якимось проміжним продуктом - головне, щоб він був креативним, а його завдання – спрощувати людям будь-які дії в їх повсякденному житті, або допомагати у веденні власного бізнесу.

В умовах сучасного Інтернету та постійно еволюціонуючих технологій, організувати власне невелике підприємство не складає жодних проблем, а сучасні соціальні мережі дозволяють знаходити інвесторів та споживачів без особливих труднощів. В таких умовах з’явилося набагато більше можливостей для розвитку свого проекту не тільки в Україні, а й за кордоном, що також призводить до росту закордонних інвестицій. Але це все при умовах, що розробка будь-якого стартапу є досить ризикованим завданням.

Більшості не вдається довести свій стартап-проект до етапу фінального тестування, не маючи мови про ринкове впровадження. За останньою статистикою, успіху (мається на увазі, вихід на ринок, але не вихід у маржинальний дохід) досягає лише 10-20% від усіх стартап-проектів.

4.1 Опис ідеї проекту

У таблиці 4.1 подано зміст ідеї стартап-проекту, можливі напрямки застосування та основні вигоди, що може отримати кінцевий користувач результату розробки, у таблиці 4.2 подано визначення сильних, слабких та нейтральних характеристик.

Таблиця 4.1 — Опис ідеї стартап-проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Створення системи підтримки прийняття рішень управлінських задач у економічно-екологічній сфері	1.Регіональне управління	Швидке опрацювання великої кількості інформації задля підтримки у прийняті рішення

Таблиця 4.2 — Визначення сильних, слабких та нейтральних характеристик

Техніко-економічні характеристики ідеї	(потенційні) товари/концепції конкурентів				W слабка сторона	N нейтральна	S сильна
	Мій проект	Poly Analyst	Neuro Shell	Brain Maker			
Вартість програмного забезпечення	Низька	Вис	Вис	Вис			+
Доступність	Низька	Вис	Вис	Сер		+	
Кроссплат.	Так	Так	Ні	Ні			+
Підтримка	+	-	+	+		+	

Отже, виходячи з попередньої таблиці можна зробити висновок, що стартап-проект є потенційно конкурентоспроможним.

4.2 Технологічний аудит ідеї проекту

За результатами аналізу таблиці 4.3 можна зробити висновок про можливість технологічної реалізації проекту.

Таблиця 4.3 — Технологічна здійсненність ідеї проекту

№п/п	Ідея проекту	Технології реалізації	Наявність технологій	Доступність технологій
	Програмний продукт для текстової аналітики	Використання інструментів компанії SAS для збору, обробки та аналізу настроїв неструктурованих даних	Наявна	Доступна
Обраною мовою програмування є SAS, використовується інструмент SAS Text Miner				

4.3 Аналіз ринкових можливостей запуску стартап-проекту

На меті є визначення ринкових можливостей, для використання під час ринкового впровадження проекту, та знаходження можливих ринкових загроз, які можуть перешкодити реалізації проекту, що дозволить спланувати напрям

розвитку проекту із урахуванням усіх цих ринкових показників, а також потреб потенційних клієнтів та пропозицій конкурентів.

У таблиці 4.4 проведено аналіз попиту: обсяг, наявність попиту, динаміка розвитку ринку.

Таблиця 4.4 — Попередня характеристика потенційного ринку стартапу

№ п/п	Показники стану ринку(найменування)	Характеристика
1	Кількість систем-конкурентів на ринку, од	4
2	Загальний обсяг продаж, грн/ум.од	75 000
3	Динаміка ринку	Постійна незмінна
4	Наявність обмежень для входу (вказати характер обмежень)	Нормативні документи державного рівня на відповідність ПЗ до законів України
5	Специфічні вимоги до стандартизації та сертифікації	Відповідно до Закону України
6	Середнє значення рентабельності в галузі(або по ринку), %	10% (відповідає середній річній ставці депозиту у гривні)

За результатами аналізу таблиці 4.4 можна зробити висновок, що ринок є привабливим для входження за попереднім оцінюванням.

Необхідно визначити потенційні групи клієнтів, та їх характеристики. Також, потрібно сформувати вимоги до кінцевого товару для кожної групи (табл. 4.5).

Таблиця 4.5 — Характеристика потенційних клієнтів стартап-проекту

Потреба, що формує ринок	Цільова аудиторія та потенційні клієнти	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
Система підтримки рішень	Регіональні, обласні державні та приватні підприємства, що керують та слідкують за економічним або екологічним станом регіону	У різних підприємств різний підхід до аналізу, висновків та підходу до постановки задач щодо контролю та регулювання станом	Простота інтерфейсу; Висока швидкість обробки великої кількості інформації; Інструкція щодо використання продукту; Технічна підтримка та супровід продукту.

Після дослідження потенційних категорій клієнтів необхідно провести дослідження ринкового середовища: сформувати таблиці факторів, що сприяють реалізації системи, та факторів, що йому заважають(табл. 4.6-4.7).

Таблиця 4.6 — Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Високий поріг потреб ринку у сфері стандартизації та ліцензування	Великі втрати при ліцензуванні, тестуванні та впровадженні систем, зв’язані з юридичною стороною	Впровадити чітку систему розробки ПЗ, відповідно до відомих умов технічного завдання
2	Зміна потреб користувачів	Клієнту потрібна буде система з додатковими умовами та можливостями	Передбачити можливість розширення системи та підвищення індексу модульності системи

Таблиця 4.7 — Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Конкуренція	Відсутність аналогів на українському ринку для вітчизняного користувача	Адаптація системи до особливостей українського ринку
2	Поява альтернативних методів моделювання	Нові методи моделювання, більш легкі в освоєнні	Розширення можливостей, максимальне спрощення

Також проведено дослідження пропозиції: було визначено загальні характеристики конкуренції на ринку (таблиця 4.8).

Таблиця 4.8 — Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому виражена дана характеристика	Вплив на діяльність компанії (можливі дії для підвищення конкурентоспроможності)
1.Вказати тип конкуренції – монополія	На ринку присутні декілька постачальників-конкурентів, але їх товар дещо відрізняється від нашої системи.	Підтримка якості системи, безперервний розвиток, покращення, вдосконалення, оновлення та підтримка.
Особливості конкурентного середовища	В чому виражена дана характеристика	Вплив на діяльність компанії (можливі дії для підвищення конкурентоспроможності)
2. За рівнем конкурентної боротьби-міжнародний	Компанії-конкуренти з інших країн	Створити основу системи підтримки рішень, щоб можна було легко локалізувати її для використання у інших країнах
3.Загалузевою ознакою-внутрішньогалузева	Система може застосовуватися в одній галузі, але її різних сферах.	Постійне вдосконалення системи текстової аналітики, що не має прив’язки до сфери
4. Конкуренція за видами товарів: -товарно-видова	Конкуренція між видами систем та методів аналітики, їх особливостями	Створити систему аналітики, враховуючи недоліки інших систем
5. За характером конкурентних переваг: -цінова	Покращення процесу створення програмного продукту, мінімізація витрат на оновлення, застосування безперервної інтеграції	Використання відкритих технологій для побудови системи
6. За інтенсивністю - не марочна	Бренд присутній, але його роль незначна	Реклама, участь у конференціях, семінарах, виставках

Після аналізу конкуренції необхідно провести детальний аналіз відносних умов конкуренції в галузі (за моделлю 5 сил М. Портера) (табл. 4.9).

Таблиця 4.9 — Аналіз конкуренції в галузі за М. Портером

Складові аналізу	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
	Poly Analyst	Наявність вже Існуючих рішень	-	Якість системи та її підтримка оновлення	Більш відомий розробник, що підтримує свою систему
Висновки:	На даний момент немає конкурентів на українському ринку	Виход на український ринок буде легшим через відсутність конкуренції	-	Вимоги клієнтів такі, як зручний інтерфейс, якість програмного продукту	Випустити систему, що буду не гірше, ніж у конкурента, але мати кращу точність прогнозування.

Виходячи с висновків аналізу конкуренції (табл. 4.9), а також з урахуванням характеристик ідеї проекту (табл. 4.2), вимог споживачів до товару (табл. 4.5) та факторів маркетингового середовища (табл.№№4.6-4.7) можна визначити та обґрунтувати перелік факторів конкурентоспроможності (табл. 4.10).

Таблиця 4.10 — Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування чинників, що роблять фактор для порівняння конкурентних проектів значущим
1	Ціна	Дешевше рішення збільшить кількість потенційних клієнтів
2	Функціональність програмного забезпечення	Велика кількість можливостей системи забезпечить перевагу перед конкурентами
3	Підтримка при використанні після покупки	Супровід та розвиток готового продукту збільшить довіру клієнтів до продукту

Після визначення факторів конкурентоспроможності (табл. 4.10) можемо провести аналіз сильних та слабких сторін стартап-проекту (табл. 4.11).

Таблиця 4.11 — Порівняльний аналіз сильних та слабких сторін проекту

№ п/п	Фактор конкурентоспроможності	Бали 1-20	Рейтинг систем-конкурентів у порівнянні з розробленою системою прогнозування						
			-3	-2	-1	0	+1	+2	+3
1	Ціна	10			*				
2	Функціональність програмного забезпечення	15		*					
3	Підтримка при використанні після покупки	9					*		

Останнім кроком маркетингового дослідження можливостей при реалізації системи підтримки прийняття рішень як стартап-проекту є побудова SWOT-аналізу (матриці сильних (Strength) та слабких (Weak) сторін, загроз (Troubles) та можливостей (Opportunities) (таблиця 4.12) на основі описаних раніше конкурентних та маркетингових загроз та можливостей, а також сильних і слабких сторін (таблиця 4.11). Список маркетингових загроз та

можливостей було складено на основі дослідження факторів загроз та факторів можливостей ринкової ситуації. Маркетингові загрози та можливості є наслідками (прогнозованими результатами) впливу ринкових факторів.

Таблиця 4.12 — SWOT-аналіз стартап-проекту

Сильні сторони:	Слабкі сторони:
Ціна, Функціональність програмного забезпечення, Підтримка при використанні	Складність розповсюджувати продукцію за кордоном.
Можливості:	Загрози:
Відсутність конкуренції на українському ринку	Зміна основних потреб клієнтів, при відсутності конкуренції необхідно підтримувати інтерес аудиторії до продукту

На основі SWOT-аналізу можемо визначити альтернативи ринкової поведінки (перелік заходів) для виведення стартап-проекту на ринок та орієнтовний оптимальний час їх ринкової реалізації з огляду на потенційні проекти конкурентів, що можуть бути виведені на ринок (табл. 4.13).

Таблиця 4.13 — Альтернативи ринкового впровадження стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Приблизна ймовірність отримання ресурсів	Приблизні строки реалізації
1	Безкоштовне розповсюдження обмеженої версії створеного програмного продукту	65%	12 місяців
2	Створення програмної системи з більш універсальним методом аналізу з подальшим платним розповсюдженням (продаж платної ліцензії)	55%	12 місяців

4.4 Розроблення ринкової стратегії проекту

Розроблення ринкової стратегії першим кроком передбачає визначення стратегії охоплення ринку: опис цільових груп потенційних споживачів (табл. 4.14).

Таблиця 4.14 — Вибір цільових груп потенційних споживачів

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Державні підприємства, що виконують нагляд та керування у економіко-екологічних сферах	Готові	Необхідно	Висока	Середня
2	Приватні підприємства, що зацікавлені у питаннях економіко-екологічної обстановці	Готові	Зацікавленість	Висока	Середня

Визначена цільова група клієнтів: Державні підприємства, що виконують нагляд та керування у економіко-екологічних сферах

Для роботи в обраних сегментах ринку сформуємо базову стратегію розвитку (табл. 4.15).

Таблиця 4.15 — Визначення базової стратегії розвитку

Обрана альтернатива розвитку стартап-проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку
Безкоштовне розповсюдження обмеженої версії створеного програмного продукту	Визначити потреби сучасного ринку для кожної з груп.	Цінова політика, функціональність продукту	Стратегія диференціації

Наступним кроком необхідно обрати стратегію конкурентної поведінки (таблиця 4.16).

Таблиця 4.16 — Визначення базової стратегії конкурентної поведінки

Чи є проект першою подібною системою на ринку?	Чи буде розробник системи шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде розробник копіювати властивості товару конкурента	Стратегія конкурентної поведінки
Ні	Шукати нових	Ні	Заняття конкурентної ніші

Також необхідно сформувати ринкову позицію, за якою споживачі будуть ідентифікувати проект(табл. 4.17).

Таблиця 4.17 — Визначення стратегії позиціонування

Вимоги до системи з боку потенційних клієнтів	Основна стратегія розвитку	Ключові конкуренто спроможні позиції власного стартап-проекту	Вибір основних асоціацій
Висока швидкість обробки інформації, Докладне керівництво для користувача, Зручний інтерфейс, Технічна підтримка користувачів	Стратегія диференціації	Позиція на основі порівняння стартапу з пропозиціями конкурентів; Відмінні особливості споживача	Автоматизація процесів; Зручність застосування; Швидкість роботи; Технічна підтримка

Виходячи з результатів дослідження, була сформована система рішень щодо поведінки стартап-компанії на ринку, яка визначає напрями роботи стартап- компанії на українському та міжнародному ринках.

4.5 Розроблення маркетингової програми стартап-проекту

У табл. 4.18 підсумуємо результати попереднього аналізу конкуренто-спроможності товару.

Таблиця 4.18 — Визначення ключових переваг концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами
1	Швидка обробка даних	Продукт має високу швидкість обробки великої кількості інформації	Швидкість обробки інформації є однією з ключових переваг продукту, що гарантує інтерес клієнтів
2	Технічна підтримка продукту	Після покупки версії продукту в процесі користування клієнту надається технічна підтримка по використанню продукту	Підтримка продукту надає клієнту впевненість у вирішенні будь-яких питань, що можуть виникнути по мірі використання продукту
3	Зменшення кількості часу на обробку інформації робочим персоналом	При використанні продукту у персоналу звільнюється час на інші робочі питання, що призводить до збільшення ефективності роботи	Оптимізація робочого часу персоналу та зменшення рутинної роботи якісно поліпшує умови роботи клієнта, що може принести додатковий прибуток.

Також була розроблена трирівнева маркетингова модель товару(табл. 4.19).

Таблиця 4.19 — Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові
1. Товар за задумом	Система для підтримки рішень у еколого-економічній сфері. Повинен бути швидким, зручним та зрозумілим
2. Товар у реальному виконанні	Властивості/характеристики
	1.Швидка обробка даних 2.Текстова аналітика 3.Багатофункціональність
	Якість: Проходження тестування
	Пакування: Відсутнє
	Марка: Відсутня
3. Товар із підкріпленням	Наявне після продажу: Технічна підтримка, навчання персоналу

На даному етапі необхідно визначити цінові межі, якими необхідно керуватись при встановленні ціни на систему, яка включає аналіз ціни на товари-аналоги або товари субститути, а також аналіз рівня доходів цільової категорії клієнтів (таблиця 4.20).

Таблиця 4.20 — Визначення меж встановлення ціни

№ п/п	Приблизна вилка вартості товарів-замінників	Приблизна вилка вартості товарів-аналогів	Приблизний рівень доходів цільової групи клієнтів	Верхня та нижня межі вартості системи
1	1000-5000$	800-4000$	10000$+	800-3000$

Після цього необхідно визначити оптимальну систему збуту, за допомогою якої буде розповсюджуватися система як сатрап-проект.(таблиця 4.21)

Таблиця 4.21 — Формування системи збуту

№ п/п	Особливості ринкової поведінки потенційних клієнтів	Функції збуту, які має виконувати постачальник системи	Глибина каналу збуту	Оптимальний канал збуту
1	Цільові клієнти – компанії, яким необхідний функціональний продукт для підтримки рішень на основі даних, що оброблює система	Побудова прямих контактів із потенційними клієнтами і їх підтримка. Формування попиту і стимулювання продажів.	Один (від виробника одразу споживачу)	Державні тендери на забезпечення технічних завдань державних підприємств; Прямі поставки підприємствам.

Фінальною складовою маркетингової програми є визначення концепції маркетингових комунікацій, що спирається на попередньо обрану основу для позиціонування, визначену специфіку поведінки клієнтів (табл. 4.22).

Таблиця 4.22 – Концепція маркетингових комунікацій

Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
Цільові клієнти: Державні та приватні підприємст-ва, що виконують контрольно-керуючу функцію у економіко-екологічній сфері окремої територіальної одиниці.	Конференції, Видання у професійних видавництвах, Новини у сфері інформаційних технологій	Позиція на основі порівняння власного продукту з продуктами конкурентів Відмінні особливості споживачів	Проінформу-вати про новий продукт та його переваги; Доказати функціональні переваги власного продукту над іншими; Збільшити потік клієнтів	Оптимізуємо робочий час робітників, Швидко оброблюємо інформацію, та надаємо технічну підтримку

4.6 Висновки до розділу 4

В даному розділі було проведено аналіз здатності успішного виведення на ринок стартап-проекту на основі системи, що була розроблена в рамках магістерської дисертації.

В процесі цього аналізу було розроблено опис самої ідеї проекту, визначено загальні напрями використання продукту, проаналізовано ринкові можливості щодо впровадження стартап-проекту, визначено характерні відмінності від конкурентів та розроблено ймовірну стратегію виходу на ринок.

Узагальнюючи проведений аналіз, можна зазначити, що проект має можливість ринкової комерціалізації проекту. Наявний постійний попит, ринок знаходиться у підвішеному стані через малу конкуренцію. З огляду на потенційні групи клієнтів, а саме на державні та приватні підприємства, та високий рівень конкурентоспроможності проекту, є чималі перспективи для впровадження стартап-проекту. Отже, подальший розвиток проекту є доцільним.

ВИСНОВКИ

Складність створення аналітичного інструментарію для розв’язання задач підтримки прийняття рішень в галузі урядування перш за все пов’язана з необхідністю урахування значної кількості кількісних та якісних факторів. Зокрема, особливостей життєвого циклу різних рівнів соціально-економічної системи держави, її структуру, зв'язки, суспільно-політичні чинники, екологічні фактори, зовнішні та внутрішні впливи різного характеру, а також передбачити можливі сценарії розвитку подій, наслідки та потенційні ризики помилок під час прийняття управлінських рішень. Саме тому питання розробки сучасного аналітичного програмного забезпечення, призначеного для використання в органах місцевого самоврядування є актуальною задачею.

В даній магістерській дисертації було розроблено та протестовано прикладну систему аналізу та категорізації еколого-економічних даних, призначену для використання в управілння територіальними громадами.

В першому розділі розглянуто інструменти SAS для інтелектуального аналізу даних. По результатам дослідження було обрано систему SAS Text Miner.

У другому розділі роботи було описано основні кроки в задачі інтелектуального аналізу даних. На першому кроці відбувається попередня обробка словника - це визначення значущих термінів: видалення стоп-слів, стемінг (визначення коренів слів) і визначення ваги термінів. На другому кроці набір даних кодується як терм-документна матриця і визначається спосіб вимірювання відстаней між документами, де відстань будується, щоб вказати, як вони схожі за змістом.

В третьому розділі було розроблено прикладну систему аналізу еколого-економічних даних за допомогою SAS Enterprise Miner. Було розроблено систему, що може категоризувати неструктуровані еколого-економічні дані, були отримано правила для категорізації. Однією з переваг даної системи є

можливість налаштувати автоматичний пошук в наявних даних, їх розподіл на різні категорії. Після обробки інформації користувач отримує не тільки результати категоризації, а також аналіз зв’язків між ключовими термінами.

Науковою новизною роботи є використанням сучасних високопродуктивних технологій накопичення та доступу даних та методів текстової аналітики у інформаційно-аналітичній системі, реалізованій у вигляді програмного додатку засобами мультиплатформного середовища аналізу даних SAS Foundation.

Запропонована інформаційно-аналітична система, спрямована перш за все на одержання соціального ефекту через покращення якості розробки бізнес-планів та інвестиційних програм, регіональних стратегій, програм соціально-економічного розвитку, стратегій розвитку громад.

ПЕРЕЛІК ПОСИЛАНЬ

1. Куць Є. С. Урбанізовані території: методологія та практика планування і управління. Мелітополь : НДІП містобудування, 2003. 219 с.
2. Варенко В. М., Братусь І. В., Дорошенко В. С., Смольніков Ю. Б., Юрченко В. О. Системний аналіз інформаційних процесів: навч. посіб. К.: Університет "Україна", 2013. 203 с.
3. Milward D. What is Text Mining, Text Analytics and Natural Language Processing?, 2018. URL: https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing
4. Mehl E., Matthias R., Quantitative Text Analysis. Handbook of multimethod measurement in psychology. NJ: Wiley, 2006. 141 p.
5. Барсегян А. А., Куприянов М. С., Холод И. И., Тесс М. Д., Елизаров С. И. Анализ данных и процессов: учеб. пособие 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.
6. Shaidah J., Hejab M. Techniques, Applications and Challenging Issue in Text Mining. *Journal of Computer Science*. 2012. Vol. 9, No. 2. P. 23.
7. Ding C., Peng H. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*. 2005. Vol. 3, No. 2. P. 185–205.
8. Zhao Y., Analysing twitter data with text mining and social network analysis, *11th Australasian Data Mining and Analytics Conference*. 2013. P. 23.
9. Murali P. Text analytics using SAS Text Miner: course notes. NC.: SAS Institute, 2014. 218 p.
10. Sethi A.Text Analytics with SAS®: Special Collection. Cary, NC: SAS Institute Inc, 2019. 568 p.
11. Murali P. Text Mining and Analysis – Practical Methods, Examples & Case Studies using SAS®.NC: SAS Institute Inc, 2013. 311 p.

12. Фирсова И.В. Социально-экономическое развитие региона и формирование нового типа потребителя. Харьков: Еспада, 2003. 20 с.
13. Хвесик М.А., Горбач Л.М., Вишневская Н., Хвесик Ю.М. Стратегия социально- экономического развития региона: монография. М.: Кондор, 2004. 376 с.
14. Терентьєв О.М., Просянкіна-Жарова Т.І., Савастьянов В.В. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України. *Реєстрація, зберігання і обробка даних.* 2016. Т. 18, № 3. С.75-86. URL: http://nbuv.gov.ua/UJRN/rzod_2016_18_3_10
15. Porter M. Algorithm for suffix striping. Program. Cambridge: CUP, 1980. 316 p.
16. Sebastiani, F. Machine learning in automated text categorization. NY: LiveLib, 2002. 135 p.
17. Martinez A. A framework for the representation of semantics. Richmond:ABK, 2002. 108 p.
18. Berry M., Browne M. Understanding Search Engines:Mathematical Modeling and Text Retrieval(Software, Environments, Tools).NY:SIAM, 2005. 205 p.
19. Duda R., Hart P., Stork D. Pattern Classification, San-Francisco: Freeman, 2000. 318 p.
20. Berrar D., Dubitzky W., Granzow M. Singular value decomposition and principal component analysis In A Practical Approach to Microarray Data Analysis, FL: BN Publishing, 2003. 109 p.
21. Deerwester S., Dumais S., Furnas G. Landauer T. Indexing by latent semantic analysis. *Journal of the Am.Soc. for Information Science*. 1990. Vol. 41, No. 6. P. 391–407.
22. Sneath P., Sokal R. Numerical taxonomy: The principles and practices of numerical classification. San-Francisco: Freeman, 1973. 573 p.

23. Rokach L. The Data Mining and Knowledge Discovery Handboo. NY: Dover, 2005. 1279 p.
24. Sholom M. Text minig. Predictive methods of analyzing unstructured information. Chicago: Anchor, 2004. 236 p.
25. Терентьєв О. М. Курс лекцій з Методів і технологій аналізу текстової інформації. Київ:НТУУ “КПІ”, 2020. 105 с.
26. Курс лекцій з SPSS. Тема 9: Кластерний аналізу. НАФИ, Москва, 2017. 46 с. URL: https://nafi.ru/upload/spss/Lection_9.pdf

ДОДАТКИ

ДОДАТОК А

ЛІСТИНГ ПРОГРАММИ

```
/* ---------- */
*------------------------------------------------------------*;
* TextParsing: Creating EM5BATCH data sets;
*------------------------------------------------------------*;
%let EM_ACTION = run;
%let EM_DEBUG =;
*------------------------------------------------------------*;
* Create workspace data set;
*------------------------------------------------------------*;
data workspace;
length property $64 value $100;
property= 'PROJECTLOCATION';
value= "C:\Workshop\!Projects";
output;
property= 'PROJECTNAME';
value= "valentin";
output;
property= 'WORKSPACENAME';
value= "EMWS1";
output;
property= 'SUMMARYDATASET';
value= 'summary';
output;
property= 'NUMTASKS';
value= 'SINGLE';
output;
property= 'EDITMODE';
value= 'M';
output;
property= 'DEBUG';
value= "&&EM_DEBUG";
```

```
output;
property= 'UNLOCK';
value= 'N';
output;
property= 'FORCERUN';
value= 'Y';
output;
run;
*------------------------------------------------------------*;
* Create actions data set:;
*------------------------------------------------------------*;
%macro emaction;
%let actionstring = %upcase(&EM_ACTION);
%if %index(&actionstring, RUN) or %index(&actionstring, REPORT) %then %do;
data actions;
length id $12 action $40;
id="TextParsing";
%if %index(&actionstring, RUN) %then %do;
action='run';
output;
%end;
%if %index(&actionstring, REPORT) %then %do;
action='report';
output;
%end;
run;
%end;
%mend;
%emaction;
*------------------------------------------------------------*;
* Execute the actions;
*------------------------------------------------------------*;
%em5batch(execute, workspace=workspace, action=actions);
```

```
/* -------------------------- */
if upcase(NAME) = "_DOCUMENT_" then do;
ROLE = "ID";
LEVEL = "NOMINAL";
end;

/* -------------------------- */
%macro main();
 %if %upcase(&EM_ACTION) eq CREATE %then %do;
  filename temp catalog 'sashelp.emtxtext.parse_create.source';
  %include temp;
  %create();
 %end;
 %if %upcase(&EM_ACTION) eq TRAIN %then %do;
  filename temp catalog 'sashelp.emtxtext.parse_train.source';
  %include temp;
  %train();
 %end;
 %if %upcase(&EM_ACTION) eq REPORT %then %do;
  filename temp catalog 'sashelp.emtxtext.parse_report.source';
  %include temp;
  %report();
 %end;
 %if %upcase(&EM_ACTION) eq SCORE %then %do;
  filename temp catalog 'sashelp.emtxtext.parse_score.source';
  %include temp;
  %score();
 %end;
  %if %upcase(&EM_ACTION) eq OPENTABLE1 %then %do;
    filename temp catalog 'sashelp.emtxtext.parse_actions.source';
    %include temp;
    filename temp;
    %openTable1;
  %end;
```

```
%mend main;


%main();


/* ---------------------------- */
*-------------------------------------------------------------*;
* TextRule: Creating EM5BATCH data sets;
*-------------------------------------------------------------*;
%let EM_ACTION = run;
%let EM_DEBUG =;
*-------------------------------------------------------------*;
* Create workspace data set;
*-------------------------------------------------------------*;
data workspace;
length property $64 value $100;
property= 'PROJECTLOCATION';
value= "C:\Workshop\!Projects";
output;
property= 'PROJECTNAME';
value= "valentin";
output;
property= 'WORKSPACENAME';
value= "EMWS1";
output;
property= 'SUMMARYDATASET';
value= 'summary';
output;
property= 'NUMTASKS';
value= 'SINGLE';
output;
property= 'EDITMODE';
value= 'M';
output;
property= 'DEBUG';
```

```
value= "&&EM_DEBUG";
output;
property= 'UNLOCK';
value= 'N';
output;
property= 'FORCERUN';
value= 'Y';
output;
run;
*------------------------------------------------------------*;
* Create actions data set:;
*------------------------------------------------------------*;
%macro emaction;
%let actionstring = %upcase(&EM_ACTION);
%if %index(&actionstring, RUN) or %index(&actionstring, REPORT) %then %do;
data actions;
length id $12 action $40;
id="TextRule";
%if %index(&actionstring, RUN) %then %do;
action='run';
output;
%end;
%if %index(&actionstring, REPORT) %then %do;
action='report';
output;
%end;
run;
%end;
%mend;
%emaction;
*------------------------------------------------------------*;
* Execute the actions;
*------------------------------------------------------------*;
%em5batch(execute, workspace=workspace, action=actions);
```

```
/* -------------------------- */
%macro main();

  %if %upcase(&EM_ACTION) eq CREATE %then %do;
    filename temp catalog 'sashelp.emtxtext.boolcat_create.source';
    %include temp;
    %create();
  %end;
  %if %upcase(&EM_ACTION) eq TRAIN %then %do;
    filename temp catalog 'sashelp.emtxtext.boolcat_train.source';
    %include temp;
    %train();
  %end;
  %if %upcase(&EM_ACTION) eq REPORT %then %do;
    filename temp catalog 'sashelp.emtxtext.boolcat_report.source';
    %include temp;
    %report();
  %end;

 %mend main;

%main();

/* -------------------------- */
%macro main;
   %if %upcase(&EM_ACTION) = CREATE %then %do;
      filename temp catalog 'sashelp.emtxtext.topic_create.source';
      %include temp;
      %create;
   %end;
   %if %upcase(&EM_ACTION) = TRAIN %then %do;
      filename temp catalog 'sashelp.emtxtext.topic_train.source';
      %include temp;
```

```
      %train;
   %end;
  %if %upcase(&EM_ACTION) = SCORE %then %do;
      filename temp catalog 'sashelp.emtxtext.topic_score.source';
      %include temp;
      %score;
   %end;
   %if %upcase(&EM_ACTION) = REPORT %then %do;
      filename temp catalog 'sashelp.emtxtext.topic_report.source';
      %include temp;
      %report;
   %end;
%mend main;

%main;

/* -------------------------- */
filename temp catalog "sashelp.emtxtext.tmt_doc_score.source";
%include temp;
filename temp catalog "sashelp.emtxtext.row_pivot_normalize.source";
%include temp;
filename temp;
filename temp catalog "sashelp.emtext.tmemclus.source";
%include temp;
filename temp catalog "sashelp.emtext.tmpred.source";
%include temp;
filename temp catalog "sashelp.emtxtext.tmc_doc_score.source";
%include temp;
filename temp catalog "sashelp.emtext.tmsort.source";
%include temp;
filename temp catalog "sashelp.emtext.tmsvd.source";
%include temp;
filename temp catalog "sashelp.emtext.tmfast.source";
%include temp;
```

```
filename temp;

libname termloc "C:\Workshop\!Projects\valentin\Workspaces\EMWS1";

/* ---------------------------- */
if upcase(NAME) = "TEXTCLUSTER_CLUSTER_" then do;
ROLE = "SEGMENT";
LEVEL = "NOMINAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB1" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB2" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB3" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB4" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB5" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB6" then do;
ROLE = "REJECTED";
end;
else
```

```
if upcase(NAME) = "TEXTCLUSTER_PROB7" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB8" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_PROB9" then do;
ROLE = "REJECTED";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD1" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD10" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD11" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD12" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD13" then do;
ROLE = "INPUT";
```

```
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD14" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD15" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD16" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD17" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD18" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD19" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD2" then do;
```

```
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD20" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD21" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD22" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD23" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD24" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD25" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
```

```
if upcase(NAME) = "TEXTCLUSTER_SVD26" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD27" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD28" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD29" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD3" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD30" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD31" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
```

```
else
if upcase(NAME) = "TEXTCLUSTER_SVD32" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD33" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD34" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD35" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD36" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD37" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD38" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
```

```
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD39" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD4" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD40" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD41" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD42" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD43" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD5" then do;
ROLE = "INPUT";
```

```
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD6" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD7" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD8" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
else
if upcase(NAME) = "TEXTCLUSTER_SVD9" then do;
ROLE = "INPUT";
LEVEL = "INTERVAL";
end;
/* ------------------------- */
%tmc_doc_score(import=&em_score_output,export=work._newexport,
termds=termloc.TextFilter2_filtterms, configds=termloc.TextCluster_tmconfig,
clusters=termloc.TextCluster_clusters, emoutstat=termloc.TextCluster_emoutstat,
_scrout=work.TextFilter2_out, svd_u=termloc.TextCluster_svd_u, svd_s=termloc.TextCluster_svd_s,
prefix=TextCluster);
data &em_score_output; set work._newexport;
/* ------------------------- */
%macro main();

  %if %upcase(&EM_ACTION) eq CREATE %then %do;
    filename temp catalog 'sashelp.emtxtext.cluster_create.source';
```

```
    %include temp;
    %create();
  %end;
  %if %upcase(&EM_ACTION) eq TRAIN %then %do;
    filename temp catalog 'sashelp.emtxtext.cluster_train.source';
    %include temp;
    %train();
  %end;
  %if %upcase(&EM_ACTION) eq REPORT %then %do;
    filename temp catalog 'sashelp.emtxtext.cluster_report.source';
    %include temp;
    %report();
  %end;
  %if %upcase(&EM_ACTION) eq SCORE %then %do;
    filename temp catalog 'sashelp.emtxtext.cluster_score.source';
    %include temp;
    %score();
  %end;
%mend main;

%main();

/* ------------------------- */
filename temp catalog "sashelp.emtext.tmemclus.source";
%include temp;
filename temp catalog "sashelp.emtext.tmpred.source";
%include temp;
filename temp catalog "sashelp.emtxtext.tmc_doc_score.source";
%include temp;
filename temp catalog "sashelp.emtext.tmsort.source";
%include temp;
filename temp catalog "sashelp.emtext.tmsvd.source";
%include temp;
filename temp catalog "sashelp.emtext.tmfast.source";
```

```
%include temp;

filename temp;


libname termloc "C:\Workshop\!Projects\valentin\Workspaces\EMWS1";



/* ------------------------- */

F_TextCluster7_cluster_ =1 ::

(OR

, (AND, (OR, "концентрації" , "концентрації" , "концентраційну" ), (NOT, "чи" )))

F_TextCluster7_cluster_ =5 ::

(OR

, (AND, (OR, "сша2" , "сша" ), (OR, "інституцій3" , "інституцій" , "інституції" , "інституції" ))

, (AND, (OR, "theo" , "them" , "then" , "theу" , "the" , "theу" , "thes" ), "рівень" )

, (AND, (NOT, (OR, "умов" , "ум" , "умов" , "умо" , "умову" )), (OR, "а." , "а" )))

F_TextCluster7_cluster_ =4 ::

(OR

, (AND, "т" , (NOT, (OR, "є." , "є" )))

, (AND, (OR, "comp" , "come" , "come" , "com" , "comf" ), (NOT, (OR, "є." , "є" )))

, (AND, (NOT, (OR, "віді" , "віду" , "відє" , "відь" , "відо" , "від" )), (OR, "компанії" , "компаніїї" , "копані" , "компанії" , "компані" , "компані" )))

F_TextCluster7_cluster_ =3 ::

(OR

, (AND, "пізніх" , (OR, "ізз" , "із" ))

, (AND, "викиди" , "радою" )

, (AND, (OR, "підгалузей" , "підгалузі" ))

, (AND, "механізму" , (OR, "кліматично" , "кліматично" , "кліматичного" , "кліматичною" , "кліматичної" , "кліматичної" ), "сфері" )

, (AND, (OR, "витриму" , "витримує" ))

, (AND, (OR, "екологічність" , "екологічніст" , "екологічність" , "екологічній" , "екологічній" ), (OR, "проживанню" , "проживання" , "промивання" , "проживанням" ))

, (AND, "механізму" , (OR, "екологогічні" , "екологічні" )))

F_TextCluster7_cluster_ =2 ::

(OR

, (AND, (NOT, (OR, "повітрю" , "повітр" , "повітря" , "повіт" , "повітро" )), (OR, "річок" , "річок" ))
```

, (AND, (OR, "лісів" , "лісів" , "лісі"), (NOT, (OR, "політиками" , "політики" , "політиками" , "політики5")))

, (AND, (OR, "on" , "on"), (NOT, (OR, "приз" , "прид" , "прин" , "пре" , "прид" , "преб" , "переть" , "прем" , "преш" , "пред" , "прев" , "прут" , "прип" , "приа" , "при" , "прем")))

, (AND, (OR, "уп" , "упп")))

Додаток Б

Term	Role	Attr	Status	Weight	Freq	# Docs
+ длить	Verb	Alpha	Keep	0.32	6288.0	139.0
+ рок	Noun	Alpha	Keep	0.30	3168.0	134.0
+ україни	Prop	Alpha	Keep	0.34	6919.0	131.0
+ стан	Noun	Alpha	Keep	0.39	2335.0	106.0
+ том	Noun	Alpha	Keep	0.34	1371.0	103.0
+ час	Noun	Alpha	Keep	0.34	1182.0	100.0
+ переть	Verb	Alpha	Keep	0.36	1868.0	92.0
+ україна	Prop	Alpha	Keep	0.43	960.0	91.0
+ яких	Noun	Alpha	Keep	0.34	965.0	91.0
+ розвитку	Noun	Alpha	Keep	0.41	3249.0	89.0
+ використання	Noun	Alpha	Keep	0.35	1651.0	86.0
Цього	Noun	Alpha	Keep	0.35	646.0	84.0
+ млн	Abbr	Alpha	Keep	0.28	1240.0	84.0
+ лихой	Adj	Alpha	Keep	0.38	841.0	83.0
+ один	Num	Alpha	Keep	0.37	1034.0	83.0
+ років	Noun	Alpha	Keep	0.33	694.0	82.0
+ система	Noun	Alpha	Keep	0.36	1448.0	81.0
+ його	Noun	Alpha	Keep	0.38	1114.0	81.0
+ може	Noun	Alpha	Keep	0.40	876.0	81.0
З	Prop	Alpha	Keep	0.38	822.0	81.0
+ території	Noun	Alpha	Keep	0.37	1069.0	80.0
+ рівні	Noun	Alpha	Keep	0.31	600.0	79.0
+ мають	Noun	Alpha	Keep	0.36	742.0	79.0
+ зокрема	Noun	Alpha	Keep	0.38	931.0	78.0
+ тис	Noun	Alpha	Keep	0.38	1634.0	78.0
+ довкілля	Noun	Alpha	Keep	0.32	1149.0	77.0
+ ніж	Noun	Alpha	Keep	0.32	485.0	77.0
+ кількість	Noun	Alpha	Keep	0.37	678.0	76.0
+ виробництва	Noun	Alpha	Keep	0.40	1076.0	76.0
+ році	Noun	Alpha	Keep	0.43	1051.0	75.0
+ такой	Adj	Alpha	Keep	0.41	719.0	74.0
+ проблема	Noun	Alpha	Keep	0.40	1031.0	74.0
Вже	Noun	Alpha	Keep	0.34	352.0	74.0
системи	Noun	Alpha	Keep	0.36	1673.0	74.0
Всіх	Noun	Alpha	Keep	0.34	483.0	74.0
+ природних	Noun	Alpha	Keep	0.38	1150.0	73.0
+ ресурсів	Noun	Alpha	Keep	0.37	1307.0	73.0
Рівень	Noun	Alpha	Keep	0.38	683.0	72.0
+ країни	Noun	Alpha	Keep	0.40	479.0	71.0
Рівень	Noun	Alpha	Keep	0.38	683.0	72.0
+ країни	Noun	Alpha	Keep	0.40	479.0	71.0
+ господарства	Noun	Alpha	Keep	0.33	560.0	70.0
+ можуть	Noun	Alpha	Keep	0.41	572.0	70.0
+ видів	Noun	Alpha	Keep	0.36	809.0	70.0

+ охорони	Noun	Alpha	Keep	0.38	1299.0	70.0
+ заходів	Noun	Alpha	Keep	0.37	1177.0	70.0
+ числі	Noun	Alpha	Keep	0.34	523.0	70.0
+ майже	Noun	Alpha	Keep	0.35	322.0	69.0
+ управління	Noun	Alpha	Keep	0.35	1388.0	69.0
Чи	Noun	Alpha	Keep	0.38	769.0	69.0
+ рівня	Noun	Alpha	Keep	0.39	872.0	68.0
+ впливу	Noun	Alpha	Keep	0.38	817.0	68.0
+ вода	Noun	Alpha	Keep	0.39	1227.0	68.0
+ зміни	Noun	Alpha	Keep	0.36	755.0	67.0
Можна	Noun	Alpha	Keep	0.43	785.0	67.0
+ середовища	Noun	Alpha	Keep	0.42	1595.0	67.0
Також	Prop	Alpha	Keep	0.29	208.0	67.0
+ буда	Noun	Alpha	Keep	0.31	372.0	67.0
+ рік	Noun	Alpha	Keep	0.38	753.0	67.0
+ вонь	Noun	Alpha	Keep	0.41	565.0	67.0
+ природный	Adj	Alpha	Keep	0.39	1351.0	67.0
+ контроль	Noun	Alpha	Keep	0.34	838.0	66.0
+ створення	Noun	Alpha	Keep	0.38	927.0	66.0
+ основних	Noun	Alpha	Keep	0.37	527.0	66.0
+ близько	Noun	Alpha	Keep	0.34	393.0	66.0
+ про	Prop	Alpha	Keep	0.38	1133.0	66.0
+ діяльності	Noun	Alpha	Keep	0.40	1350.0	66.0
+ повітря	Noun	Alpha	Keep	0.33	1084.0	65.0
+ понад	Noun	Alpha	Keep	0.38	463.0	65.0
+ водить	Verb	Alpha	Keep	0.40	1078.0	65.0
+ захисту	Noun	Alpha	Keep	0.37	629.0	65.0
+ відповідно	Noun	Alpha	Keep	0.36	741.0	65.0
+ забезпечення	Noun	Alpha	Keep	0.38	1306.0	65.0
роботи	Noun	Alpha	Keep	0.34	486.0	65.0

Додаток В